PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C07H 21/04, C12N 15/00, 1/20, 9/14, A61K 38/46

A1

(11) International Publication Number: WO 97/33903

(43) International Publication Date: 18 September 1997 (18.09.97)

(21) International Application Number: PCT/US96/03239

(22) International Filing Date: 11 March 1996 (11.03.96)

(71) Applicants *(for all designated States except US)*: SMITHKLINE BEECHAM CORPORATION [US/US]; Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US). HUMAN GENOME SCIENCES, INC. [US/US]; 9410 Key West Avenue, Rockville, MD 20850-3338 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: WEI, Ying-Fei [CN/US]; 13524 Straw Bale Road, Darnestown, MD 20878 (US).

(74) Agents: GIMMI, Edward, R. et al.; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US).

(81) Designated States: AL, AM, AU, BB, BG, BR, CA, CN, CZ, EE, FI, GE, HU, IS, JP, KG, KP, KR, LK, LR, LT, LV, MD, MG, MK, MN, MX, NO, NZ, PL, RO, SG, SI, SK, TR, TT, UA, US, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

**Published**
*With international search report.*

(54) Title: HUMAN MutY

(57) Abstract

A human MutY polypeptide and DNA (RNA) encoding such polypeptide and a procedure for producing such polypeptide by recombinant techniques is disclosed. Also disclosed are methods for utilizing such polypeptide for preventing and/or treating diseases associated with a mutation in this gene. Diagnostic assays for identifying mutations in nucleic acid sequence encoding a polypeptide of the present invention and for detecting altered levels of the polypeptide of the present invention for detecting diseases, for example, cancer, are also disclosed.

*FOR THE PURPOSES OF INFORMATION ONLY*

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

| | | | | | |
|---|---|---|---|---|---|
| AM | Armenia | GB | United Kingdom | MW | Malawi |
| AT | Austria | GE | Georgia | MX | Mexico |
| AU | Australia | GN | Guinea | NE | Niger |
| BB | Barbados | GR | Greece | NL | Netherlands |
| BE | Belgium | HU | Hungary | NO | Norway |
| BF | Burkina Faso | IE | Ireland | NZ | New Zealand |
| BG | Bulgaria | IT | Italy | PL | Poland |
| BJ | Benin | JP | Japan | PT | Portugal |
| BR | Brazil | KE | Kenya | RO | Romania |
| BY | Belarus | KG | Kyrgystan | RU | Russian Federation |
| CA | Canada | KP | Democratic People's Republic of Korea | SD | Sudan |
| CF | Central African Republic | | | SE | Sweden |
| CG | Congo | KR | Republic of Korea | SG | Singapore |
| CH | Switzerland | KZ | Kazakhstan | SI | Slovenia |
| CI | Côte d'Ivoire | LI | Liechtenstein | SK | Slovakia |
| CM | Cameroon | LK | Sri Lanka | SN | Senegal |
| CN | China | LR | Liberia | SZ | Swaziland |
| CS | Czechoslovakia | LT | Lithuania | TD | Chad |
| CZ | Czech Republic | LU | Luxembourg | TG | Togo |
| DE | Germany | LV | Latvia | TJ | Tajikistan |
| DK | Denmark | MC | Monaco | TT | Trinidad and Tobago |
| EE | Estonia | MD | Republic of Moldova | UA | Ukraine |
| ES | Spain | MG | Madagascar | UG | Uganda |
| FI | Finland | ML | Mali | US | United States of America |
| FR | France | MN | Mongolia | UZ | Uzbekistan |
| GA | Gabon | MR | Mauritania | VN | Viet Nam |

# HUMAN MutY

## BACKGROUND OF THE INVENTION

The GO system includes 7,8-dihydro-8-oxoguanine, the structure of the predominant tautomeric form of the GO lesion. Oxidative damage can lead to GO lesions in DNA. MutY removes the misincorporated adenine from the A/GO mispairs that result from error-prone replication past the GO lesion. Repair polymerases are much less error-prone during trans lesion synthesis and can lead to a C/GO pair. Oxidative damage can also lead to 8-oxo-dGTP. Inaccurate replication could result in the misincorporation of 8-oxo-dGTP opposite template A residues, leading to A/GO mispairs. MutY could be involved in the mutation process because it is active on the A/GO substrate and would remove the template A, leading to the AT → CG transversions that are characteristic of a MutT strain. The 8-oxo-dGTP could also be incorporated opposite template cytosines, resulting in a damaged C/GO pair that could be corrected by MutM.

This invention relates to newly identified polynucleotides, polypeptides encoded by such polynucleotides, the use of such polynucleotides and polypeptides, as well as the production of such polynucleotides and polypeptides. More particularly, the polypeptide of the present invention has been putatively identified as a human homologue of the *E. coli* MutY gene, sometimes hereinafter referred to as "hMYH".

Mismatches arise in DNA through DNA replication errors, through DNA recombination, and following exposure of DNA to deaminating or oxidating environments. Cells have a host of strategies that counter the threat to their genetic integrity from mismatched and chemically damaged base pairs (Friedberg, EC, DNA repair, W.H. Freeman, New York (1985)). With regard specifically to mismatch repair of replication errors, *Escherichia coli* and *Salmonella typhimurium* direct the repair to the unmethylated newly synthesized DNA strand by *dam* methylation at d(GATC)

sequences, using the MutHLS systems (Clavery, J.P. and Lacks, S.A., Microbiol., **Rev.** **50**:133-165 (1986); Modrich, P. Annu. Rev. Genet. **25**:229-253 (1991); Radman, M. and R. Wagner, Annu. Rev. Genet. **20**:523-528 (1986)). The very short patch pathway of *E. coli* is specific for a correction of T/G mismatches (a mismatch indicated by a slash) and is responsible for the correction of deaminated 5-methylcytosine (Jones, M., et al., Genetics, **115**:605-610 (1987); Lieb, M., Mod. Gen. Genet. **181**:118-125 (1983); Lieb, M., and D. Read, Genetics **114**:1041-1060 (1986); Raposa, S. and N.S. Fox, Genetics **117**:381-390 (1987)).

The *E. coli* MutY pathway corrects A/G and A/C mismatches, as well as adenines paired with 7,8-dihydro-8-oxo-deoxyguanine (8-oxoG or GO) (Au, K.G., *et al.*, Proc. Natl. Acad. Sci. USA **85**:9163-9166 (1988); Lu, A.L. and D.Y. Chang, Genetics, **118**:593-600 (1988); Michaels, M.L., *et al.*, Proc. Natl. Acad. Sci. USA, **89**:7022-7025 (1992); Michaels, M.L., *et al.*, Biochemistry, **31**:10964-10968 (1992); Radicella, J.P., *et al.*, Proc. Natl. Acad. Sci. USA, **85**:9674-9678 (1988); Su, S.-S., *et al.*, J. Biol. Chem. **263**:6829-6835 1988). The 39-kDa MutY protein shares some homology with *E. coli* endonuclease III and contains a $[4Fe\text{-}4S]^{2+}$ cluster (Lu, A.-L., *et al.*, 1994, Unpublished data; Michaels, M.L., *et al.*, Nucleic Acids Res. **18**:3841-3845 (1990); Tsai-Wu, J.-J., *et al.*, Proc, Natl. Acad. Sci. USA **89**:8779-8783 (1992); Tsai-Wu, J.-J., *et al.*, J. Bacteriol. **173**:1902-1910 (1991)). The MutY preparation of Tsai-Wu *et al.* (Tsai-Wu, J.-J., *et al.*, Proc. Natl. Acad. Sci. USA **89**:8779-8783 (1992)) has both DNA *N*-glycosylase and apurinic or apyrimidinic (AP) endonuclease activities, whereas those purified by Au *et al.* (Au K.G., *et al.*, Proc. Natl. Acad. Sci. USA, **86**:8877-8881 (1989), and Michaels *et al.* (Michaels, M.L., *et al.*, Proc. Natl. Acad. Sci. USA, **89**:7022-7025 (1992); Michaels, M.L., *et al.*, Biochemistry, **31**:10964-10968 (1992) possess only the glycosylase activity. DNA glycosylase specifically excises the mispaired adenine from the mismatch and the AP endonuclease cleaves the first phosphodiester bond 3′ to the resultant AP site (Au K.G., *et al.*, Proc. Natl. Acad. Sci. USA, **86**:8877-8881 (1989); Tsai-Wu, J.-J., *et al.*, Proc, Natl. Acad. Sci. USA **89**:8779-8783 (1992).

Repair by the MutY pathway involves a short repair tract and DNA polymerase I (Radicella, J.P., *et al.*, J. Bacteriol., **175**:7732-7736 (1993); Tsai-Wu, J.-J., and A.-L. Lu, Mol. Gen. Genet. **244**:444-450 (1994)).

The mismatch repair strategy detailed above has been evolutionarily conserved. Genetic analysis suggests that *Saccharomyces cerevisiae* has a repair system analogous to the bacterial *dam* methylation-dependent pathway (Bishop, D.K., *et al.*, Nature (London) **243**:362-364 (1987); Reenan, R.A. and R.D. Kolodner, Genetics, **132**:963-973 (1992); Reenan, R.A. and R.D. Kolodner, Genetics, **132**:975-985 (1992); Williamson, M., *et al.*, Genetics, **110**:609-646 (1985)). This pathway is functionally homologous to the *E. coli* very sort patch pathway for the correction of deaminated 5-methylcytosine.

Two mutator genes in *E. coli*, the *mutY* and the *mutM* genes (Cabrera *et al.*, J. Bacteriol., 170:5405-5407 (1988); and Nghiem, Y., et al., PNAS, USA, 85:9163-9166 (1988)) have been described, which work together to prevent mutations from certain types of oxidative damage, dealing in particular with the oxidized guanine lesion, 8-oxodGuanine (Michaels *et al.*, PNAS, USA, 89:7022-7025 (1992). In Michaels, M.L., and Miller, J.H., J. Bacteriol., 174:6321-6325 (1992) is a summary of the concerted action of these two enzymes, both of which are glycosylases. The MutM protein removes 8-oxodG from the DNA, and the resulting AP site is repaired to restore the G:C base pair. Some lesions are not repaired before replication, which results in 50% insertion of an A across from the 8-oxodG, which can lead to a G:C to T:A transversion at the next round of replication. However, the MutY protein removes the A across from 8-oxodG and repair synthesis restore a C most of the time, allowing the MutM protein another opportunity to repair the lesion. In accordance with this, mutators lacking either the MutM or MutY protein have an increase specifically in the G:C to T:A transversion (Cabrera *et al.*, Id., (1988); and Nghiem, Y., et al., Id. (1988)), and cells lacking both enzymes have an enormous increase in this base substitution (Michaels *et al.*, Id. (1992). A third protein, the product of the *mutT* gene, prevents the incorporation of 8-oxodGTP by hydrolyzing the oxidized triphosphate back to the

monophosphate (Maki, H., and Sekiguchi, M., Nature, 355:273-275 (1992)), preventing A:T to C:G transversions.

Accordingly, there exists a need in the art for identification and characterization of genes and proteins which modulate the human cellular mutation rate, for use as, among other things, markers in cancer and diseases associated with DNA repair. In particular, there is a need for isolating and characterizing human mismatch repair genes and proteins, which are essential to proper development and health of tissues and organs, such as the colon, and which can, among other things, play a role in preventing, ameliorating, diagnosing or correcting dysfunctions or disease, particularly cancer, and most particularly colon cancer, such as, for example, HNPCC (non-polyposis colon cancer).

In accordance with one aspect of the present invention, there is provided a novel mature polypeptide, as well as biologically active and diagnostically or therapeutically useful fragments, analogs and derivatives thereof. The polypeptide of the present invention is of human origin.

In accordance with another aspect of the present invention, there are provided isolated nucleic acid molecules encoding a polypeptide of the present invention including mRNAs, cDNAs, genomic DNAs as well as analogs and biologically active and diagnostically or therapeutically useful fragments thereof.

In accordance with yet a further aspect of the present invention, there is provided a process for producing such polypeptide by recombinant techniques comprising culturing recombinant prokaryotic and/or eukaryotic host cells, containing a nucleic acid sequence encoding a polypeptide of the present invention, under conditions promoting expression of said protein and subsequent recovery of said protein.

In accordance with yet a further aspect of the present invention, there is provided a process for utilizing such polypeptide, or polynucleotide encoding such polypeptide, for therapeutic purposes, for example, to repair oxidative damage to DNA and prevent mutations from oxidative lesions, treat genetic diseases related to a

mutated hMYH gene, for example, xeroderma pigmentosum and neoplasia, and to diagnose an abnormal transformation of cells, particularly cancer, and most particularly colon cancer, such as for example HNPCC, and/or to diagnose a susceptibility to abnormal transformation of cells, particularly cancer, and most particularly colon cancer, such as for example HNPCC.

In accordance with yet a further aspect of the present invention, there are provided antibodies against such polypeptides.

In accordance with yet a further aspect of the present invention, there is also provided nucleic acid probes comprising nucleic acid molecules of sufficient length to specifically hybridize to a nucleic acid sequence of the present invention.

In accordance with still another aspect of the present invention, there are provided diagnostic assays for detecting diseases or susceptibility to diseases related to mutations in the nucleic acid sequences encoding a polypeptide of the present invention. In accordance with a further aspect of the invention is a process for diagnosing a cancer comprising determining from a sample derived from a patient a decreased level of activity of polypeptide having the sequence of SEQ ID NO: 2.

In accordance with a further aspect of the invention is a process for diagnosing a cancer comprising determining from a sample derived from a patient a decreased level of expression of a gene encoding a polypeptide having the sequence of SEQ ID NO: 2.

In accordance with a further aspect of the invention is a process for diagnosing a cancer comprising determining from a sample derived from a patient a decreased level of activity of polypeptide having the sequence of SEQ ID NO: 9.

In accordance with a further aspect of the invention is a process for diagnosing a cancer comprising determining from a sample derived from a patient a decreased level of expression of a gene encoding a polynucleotide having the sequence of SEQ ID NO:9.

In accordance with yet a further aspect of the present invention, there is provided a process for utilizing such polypeptides, or polynucleotides encoding such

polypeptides, for *in vitro* purposes related to scientific research, for example, synthesis of DNA and manufacture of DNA vectors.

These and other aspects of the present invention should be apparent to those skilled in the art from the teachings herein.

The following drawings are illustrative of embodiments of the invention and are not meant to limit the scope of the invention as encompassed by the claims.

Figure 1 is an illustration of the cDNA and corresponding deduced amino acid sequence of the polypeptide of the present invention. The nucleotide sequence of hMYH is shown with the numbering relative to the A of the ATG translation start site (+1). The amino acid sequence is shown below in single letter code and is also numbered in the margin.

Figure 2 is an amino acid sequence comparison between the polypeptide of the present invention (top line) and *E. coli* MutY protein (bottom line).

In accordance with an aspect of the present invention, there is provided an isolated nucleic acid (polynucleotide) which encodes for the mature polypeptide having the deduced amino acid sequence of Figure 1 (SEQ ID NO:2).

The polynucleotide of this invention may be obtained from numerous tissues of the human body, including brain and testes. The polynucleotide of this invention was discovered in a cDNA library derived from a human cerebellum. The hMYH gene contains 15 introns, and is 7.1 kb long. The 16 exons encode a nuclear protein of 535 amino acids that displays 41% identity to the *E. coli* MutY protein, which provides an important function in the repair of oxidative damaged DNA, and helps to prevent mutations from oxidative lesions. The hMYH gene maps on the short arm of chromosome 1, between p32.1 and p34.3. There is extensive homology between the hMYH protein and the *E. coli* MutY protein with extensive homology near the beginning of the *E. coli* protein, which is characterized by a string of 14 identical amino acids.

The polynucleotide of the present invention may be in the form of RNA or in the form of DNA, which DNA includes cDNA, genomic DNA, and synthetic DNA.

The DNA may be double-stranded or single-stranded, and if single stranded may be the coding strand or non-coding (anti-sense) strand. The coding sequence which encodes the mature polypeptide may be identical to the coding sequence shown in Figure 1 (SEQ ID NO:1) or may be a different coding sequence which coding sequence, as a result of the redundancy or degeneracy of the genetic code, encodes the same mature polypeptide as the DNA of Figure 1 (SEQ ID NO:1).

The polynucleotide which encodes for the mature polypeptide of Figure 1 (SEQ ID NO:2) may include, but is not limited to: only the coding sequence for the mature polypeptide; the coding sequence for the mature polypeptide and additional coding sequence such as a leader or secretory sequence or a proprotein sequence; the coding sequence for the mature polypeptide (and optionally additional coding sequence) and non-coding sequence, such as introns or non-coding sequence 5' and/or 3' of the coding sequence for the mature polypeptide.

Thus, the term "polynucleotide encoding a polypeptide" encompasses a polynucleotide which includes only coding sequence for the polypeptide as well as a polynucleotide which includes additional coding and/or non-coding sequence.

The present invention further relates to variants of the hereinabove described polynucleotides which encode for fragments, analogs and derivatives of the polypeptide having the deduced amino acid sequence of Figure 1 (SEQ ID NO:2). The variant of the polynucleotide may be a naturally occurring allelic variant of the polynucleotide or a non-naturally occurring variant of the polynucleotide.

Thus, the present invention includes polynucleotides encoding the same mature polypeptide as shown in Figure 1 (SEQ ID NO:2) as well as variants of such polynucleotides which variants encode for a fragment, derivative or analog of the polypeptide of Figure 1 (SEQ ID NO:2). Such nucleotide variants include deletion variants, substitution variants and addition or insertion variants. Certain specific variants, among other, are provided by the present invention, such as, an isolated nucleic acid having a cytosine (C) at position 366 and/or position 729 of the nucleotide sequence of Figure 1 (SEQ ID NO:1). Certain other specific variants, among other, are

provided by the present invention, such as, an isolated nucleic acid having a cytosine (C) at position 1095 of the nucleotide sequence of Figure 1 (SEQ ID NO:1). Further specific variants include, but are not limited to, an isolated polypeptide sequence having a glutamine (Q) at position 365 of the amino acid sequence in Figure 1 (SEQ ID NO:2).

As hereinabove indicated, the polynucleotide may have a coding sequence which is a naturally occurring allelic variant of the coding sequence shown in Figure 1 (SEQ ID NO:1). As known in the art, an allelic variant is an alternate form of a polynucleotide sequence which may have a substitution, deletion or addition of one or more nucleotides, which does not substantially alter the function of the encoded polypeptide.

The present invention also includes polynucleotides, wherein the coding sequence for the mature polypeptide may be fused in the same reading frame to a polynucleotide sequence which aids in expression and secretion of a polypeptide from a host cell, for example, a leader sequence which functions as a secretory sequence for controlling transport of a polypeptide from the cell. Thus, for example, the polynucleotide of the present invention may encode for a mature protein, or for a protein having a prosequence or for a protein having both a prosequence and a presequence (leader sequence).

The polynucleotides of the present invention may also have the coding sequence fused in frame to a marker sequence which allows for purification of the polypeptide of the present invention. The marker sequence may be a hexa-histidine tag supplied by a pQE vector to provide for purification of the mature polypeptide fused to the marker in the case of a bacterial host, or, for example, the marker sequence may be a hemaglutinin (HA) tag when a mammalian host, e.g. COS-7 cells, is used. The HA tag corresponds to an epitope derived from the influenza hemaglutinin protein (Wilson, I., et al., Cell, 37:767 (1984)).

The term "gene" means the segment of DNA involved in producing a polypeptide chain; it includes regions preceding and following the coding region

(leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

Fragments of the full length gene of the present invention may be used as a hybridization probe for a cDNA library to isolate the full length cDNA and to isolate other cDNAs which have a high sequence similarity to the gene or similar biological activity. Probes of this type have at least 15 bases, preferably 30 bases and most preferably 50 or more bases. The probe may also be used to identify a cDNA clone corresponding to a full length transcript and a genomic clone or clones that contain the complete gene including regulatory and promotor regions, exons, and introns. An example of a screen comprises isolating the coding region of the gene by using the known DNA sequence to synthesize an oligonucleotide probe. Labeled oligonucleotides having a sequence complementary to that of the gene of the present invention are used to screen a library of human cDNA, genomic DNA or mRNA to determine which members of the library the probe hybridizes to.

The present invention further relates to polynucleotides which hybridize to the hereinabove-described sequences if there is at least 70%, preferably at least 90%, and more preferably at least 95% identity between the sequences. The present invention particularly relates to polynucleotides which hybridize under stringent conditions to the hereinabove-described polynucleotides. As herein used, the term "stringent conditions" means hybridization will occur only if there is at least 95% and preferably at least 97% identity between the sequences. The polynucleotides which hybridize to the hereinabove described polynucleotides in a preferred embodiment encode polypeptides which either retain substantially the same biological function or activity as the mature polypeptide encoded by the cDNAs of Figure 1 (SEQ ID NO:1).

Alternatively, the polynucleotide may have at least 15 bases, preferably at least 30 bases, and more preferably at least 50 bases which hybridize to a polynucleotide of the present invention and which has an identity thereto, as hereinabove described, and which may or may not retain activity. For example, such polynucleotides may be

employed as probes for the polynucleotide of SEQ ID NO:1, for example, for recovery of the polynucleotide or as a diagnostic probe or as a PCR primer.

Thus, the present invention is directed to polynucleotides having at least a 70% identity, preferably at least 90% and more preferably at least a 95% identity to a polynucleotide which encodes the polypeptide of SEQ ID NO:2 and polynucleotides complementary thereto as well as portions thereof, which portions have at least 15 consecutive bases, preferably 30 consecutive bases and most preferably at least 50 consecutive bases and to polypeptides encoded by such polynucleotides.

The present invention further relates to a polypeptide which has the deduced amino acid sequence of Figure 1 (SEQ ID NO:2), as well as fragments, analogs and derivatives of such polypeptide.

The terms "fragment," "derivative" and "analog" when referring to the polypeptide of Figure 1 (SEQ ID NO:2) means a polypeptide which retains essentially the same biological function or activity as such polypeptide. Thus, an analog includes a proprotein which can be activated by cleavage of the proprotein portion to produce an active mature polypeptide.

The polypeptide of the present invention may be a recombinant polypeptide, a natural polypeptide or a synthetic polypeptide, preferably a recombinant polypeptide.

The fragment, derivative or analog of the polypeptide of Figure 1 (SEQ ID NO:2) may be (i) one in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code, or (ii) one in which one or more of the amino acid residues includes a substituent group, or (iii) one in which the mature polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or (iv) one in which the additional amino acids are fused to the mature polypeptide, such as a leader or secretory sequence or a sequence which is employed for purification of the mature polypeptide or a proprotein sequence.

Such fragments, derivatives and analogs are deemed to be within the scope of those skilled in the art from the teachings herein.

The polypeptides and polynucleotides of the present invention are preferably provided in an isolated form, and preferably are purified to homogeneity.

The term "isolated" means that the material is removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotides could be part of a vector and/or such polynucleotides or polypeptides could be part of a composition, and still be isolated in that such vector or composition is not part of its natural environment.

The polypeptides of the present invention include the polypeptide of SEQ ID NO:2 (in particular the mature polypeptide) as well as polypeptides which have at least 70% similarity (preferably at least 70% identity) to the polypeptide of SEQ ID NO:2 and more preferably at least 90% similarity (more preferably at least 90% identity) to the polypeptide of SEQ ID NO:2 and still more preferably at least 95% similarity (still more preferably at least 95% identity) to the polypeptide of SEQ ID NO:2 and also include portions of such polypeptides with such portion of the polypeptide generally containing at least 30 amino acids and more preferably at least 50 amino acids.

As known in the art "similarity" between two polypeptides is determined by comparing the amino acid sequence and its conserved amino acid substitutes of one polypeptide to the sequence of a second polypeptide. Moreover, also known in the art is "identity" which means the degree of sequence relatedness between two polypptide or two polynucleotides sequences as determined by the identity of the match between two strings of such sequences. Both identity and similarity can be readily calculated. While there exist a number of methods to measure identity and similarity between two polynucleotide or polypeptide sequences, the terms "identity" and "similarity" are well known to skilled artisans (Carillo, H., and Lipman, D., SIAM

J. Applied Math., 48: 1073 (1988). Methods commonly employed to determine identity or similarity between two sequences include, but are not limited to disclosed in Guide to Huge Computers, Martin J. Bishop, ed., Academic Press, San Diego, 1994, and Carillo, H., and Lipman, D., SIAM J. Applied Math., 48: 1073 (1988). Preferred methods to determine identity are designed to give the largest match between the two sequences tested. Methods to determine identity and similarity are codified in computer programs. Preferred computer program methods to determine identity and similarity between two sequences include, but are not limited to, BLASTP, BLASTN, FASTA.

Fragments or portions of the polypeptides of the present invention may be employed for producing the corresponding full-length polypeptide by peptide synthesis; therefore, the fragments may be employed as intermediates for producing the full-length polypeptides. Fragments or portions of the polynucleotides of the present invention may be used to synthesize full-length polynucleotides of the present invention.

The present invention also relates to vectors which include polynucleotides of the present invention, host cells which are genetically engineered with vectors of the invention and the production of polypeptides of the invention by recombinant techniques.

Host cells are genetically engineered (transduced or transformed or transfected) with the vectors of this invention which may be, for example, a cloning vector or an expression vector. The vector may be, for example, in the form of a plasmid, a viral particle, a phage, etc. The engineered host cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying the genes of the present invention. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan.

The polynucleotides of the present invention may be employed for producing polypeptides by recombinant techniques. Thus, for example, the polynucleotide may

be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors include chromosomal, nonchromosomal and synthetic DNA sequences, e.g., derivatives of SV40; bacterial plasmids; phage DNA; baculovirus; yeast plasmids; vectors derived from combinations of plasmids and phage DNA, viral DNA such as vaccinia, adenovirus, fowl pox virus, and pseudorabies. However, any other vector may be used as long as it is replicable and viable in the host.

The appropriate DNA sequence may be inserted into the vector by a variety of procedures. In general, the DNA sequence is inserted into an appropriate restriction endonuclease site(s) by procedures known in the art. Such procedures and others are deemed to be within the scope of those skilled in the art.

The DNA sequence in the expression vector is operatively linked to an appropriate expression control sequence(s) (promoter) to direct mRNA synthesis. As representative examples of such promoters, there may be mentioned: LTR or SV40 promoter, the E. coli. lac or trp, the phage lambda $P_L$ promoter and other promoters known to control expression of genes in prokaryotic or eukaryotic cells or their viruses. The expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression.

In addition, the expression vectors preferably contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or such as tetracycline or ampicillin resistance in E. coli.

The vector containing the appropriate DNA sequence as hereinabove described, as well as an appropriate promoter or control sequence, may be employed to transform an appropriate host to permit the host to express the protein.

As representative examples of appropriate hosts, there may be mentioned: bacterial cells, such as E. coli, Streptomyces, Salmonella typhimurium; fungal cells, such as yeast; insect cells such as Drosophila S2 and Spodoptera Sf9; animal cells such as CHO, COS or Bowes melanoma; adenoviruses; plant cells, etc. The selection of an

appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein.

More particularly, the present invention also includes recombinant constructs comprising one or more of the sequences as broadly described above. The constructs comprise a vector, such as a plasmid or viral vector, into which a sequence of the invention has been inserted, in a forward or reverse orientation. In a preferred aspect of this embodiment, the construct further comprises regulatory sequences, including, for example, a promoter, operably linked to the sequence. Large numbers of suitable vectors and promoters are known to those of skill in the art, and are commercially available. The following vectors are provided by way of example; Bacterial: pQE70, pQE60, pQE-9 (Qiagen), pBS, pD10, phagescript, psiX174, pBluescript SK, pBSKS, pNH8A, pNH16a, pNH18A, pNH46A (Stratagene); pTRC99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia); Eukaryotic: pWLNEO, pSV2CAT, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, pSVL (Pharmacia). However, any other plasmid or vector may be used as long as they are replicable and viable in the host.

Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers. Two appropriate vectors are pKK232-8 and pCM7. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda $P_R$, $P_L$ and trp. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art.

In a further embodiment, the present invention relates to host cells containing the above-described constructs. The host cell can be a higher eukaryotic cell, such as a mammalian cell, or a lower eukaryotic cell, such as a yeast cell, or the host cell can be a prokaryotic cell, such as a bacterial cell. Introduction of the construct into the host cell can be effected by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation (Davis, L., Dibner, M., Battey, I., Basic Methods in Molecular Biology, (1986)).

The constructs in host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence. Alternatively, the polypeptides of the invention can be synthetically produced by conventional peptide synthesizers.

Mature proteins can be expressed in mammalian cells, yeast, bacteria, or other cells under the control of appropriate promoters. Cell-free translation systems can also be employed to produce such proteins using RNAs derived from the DNA constructs of the present invention. Appropriate cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described by Sambrook, et al., Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor, N.Y., (1989), the disclosure of which is hereby incorporated by reference.

Transcription of the DNA encoding the polypeptides of the present invention by higher eukaryotes is increased by inserting an enhancer sequence into the vector. Enhancers are cis-acting elements of DNA, usually about from 10 to 300 bp that act on a promoter to increase its transcription. Examples include the SV40 enhancer on the late side of the replication origin bp 100 to 270, a cytomegalovirus early promoter enhancer, the polyoma enhancer on the late side of the replication origin, and adenovirus enhancers.

Generally, recombinant expression vectors will include origins of replication and selectable markers permitting transformation of the host cell, e.g., the ampicillin resistance gene of E. coli and S. cerevisiae TRP1 gene, and a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), α-factor, acid phosphatase, or heat shock proteins, among others. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably, a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium. Optionally, the heterologous sequence can encode a fusion

protein including an N-terminal identification peptide imparting desired characteristics, e.g., stabilization or simplified purification of expressed recombinant product.

Useful expression vectors for bacterial use are constructed by inserting a structural DNA sequence encoding a desired protein together with suitable translation initiation and termination signals in operable reading phase with a functional promoter. The vector will comprise one or more phenotypic selectable markers and an origin of replication to ensure maintenance of the vector and to, if desirable, provide amplification within the host. Suitable prokaryotic hosts for transformation include E. coli, Bacillus subtilis, Salmonella typhimurium and various species within the genera Pseudomonas, Streptomyces, and Staphylococcus, although others may also be employed as a matter of choice.

As a representative but nonlimiting example, useful expression vectors for bacterial use can comprise a selectable marker and bacterial origin of replication derived from commercially available plasmids comprising genetic elements of the well known cloning vector pBR322 (ATCC 37017). Such commercial vectors include, for example, pKK223-3 (Pharmacia Fine Chemicals, Uppsala, Sweden) and GEM1 (Promega Biotec, Madison, WI, USA). These pBR322 "backbone" sections are combined with an appropriate promoter and the structural sequence to be expressed.

Following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter is induced by appropriate means (e.g., temperature shift or chemical induction) and cells are cultured for an additional period.

Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification.

Microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents, such methods are well known to those skilled in the art.

Various mammalian cell culture systems can also be employed to express recombinant protein. Examples of mammalian expression systems include the COS-7

lines of monkey kidney fibroblasts, described by Gluzman, Cell, 23:175 (1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA sequences derived from the SV40 splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

The polypeptide can be recovered and purified from recombinant cell cultures by methods including ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography. Protein refolding steps can be used, as necessary, in completing configuration of the mature protein. Finally, high performance liquid chromatography (HPLC) can be employed for final purification steps.

The polypeptides of the present invention may be a naturally purified product, or a product of chemical synthetic procedures, or produced by recombinant techniques from a prokaryotic or eukaryotic host (for example, by bacterial, yeast, higher plant, insect and mammalian cells in culture). Depending upon the host employed in a recombinant production procedure, the polypeptides of the present invention may be glycosylated or may be non-glycosylated. Polypeptides of the invention may also include an initial methionine amino acid residue.

The fragments, analogs and derivatives of the polypeptides of the present invention may be assayed for determination of mismatch-nicking activity and glycosylase activity. As an example of such an assay, protein samples are incubated with 1.8 fmol of either a 5′-end-labeled 116-mer, a 3′-end-labeled 120-mer, or a 3'-end-labeled 20-mer duplex DNA containing mismatches (see Yeh, Y.-C., *et al.*, 1991, J. Bio. Chem. 266:6480-6486); (Roelen, H.C.P.F., *et al.*, 1991, Nucleic Acids Res.

19:4361-4369) in a 20 μl reaction mixture containing 10 mM Tris-HCl (pH 7.6), 5 μM $ZnCl_2$, 0.5 mM DTT, 0.5 mM EDTA, and 1.5% glycerol. Following a 2 hour incubation at 37°C, the reaction products are lyophilized and dissolved in a solution containing 3 μl of 90% (vol/vol) formamide, 10 mM EDTA, 0.1% (wt/vol) xylene cyanol, and 0.1% (wt/vol) bromophenol blue. After heating at 90°C for 3 minutes, DNA samples are analyzed on 8% polyacrylamide-8.3 M urea DNA sequencing gels (Maxam, A.M. and W. Gilbert, 1980, Methods Enzymol., 65:499-560), and the gel was then autoradiographed. The DNA glycosylase activity was monitored by adding piperidine, after the enzyme incubation, to a final concentration of 1 M. After 30 minutes of incubation at 90°C, the reaction products are analyzed as described above.

An Enzyme binding assay may also be performed wherein protein-DNA complexes are analyzed on 4% polyacrylamide gels in 50 mM Tris-borate (pH 8.3) and 1 mM EDTA. Protein samples are incubated with 3'-end-labeled 20-bp oligonucleotides as in the nicking assay, except 20 ng or poly(dI-dC) is added to each reaction mixture. Bovine serine albumin (1 μg) is added as indicated to the binding assay. For the binding competition assay, in addition to the 1.8 fmol of labeled 20-mer substrates, unlabeled 19-mer DNAs containing A/G, A/GO, or C·G pairings are added in excess of up to 180 fmol.

The invention provides a process for diagnosing a disease, particularly cancer, comprising determining from a sample derived from a patient a decreased level of activity of polypeptide having the sequence of Figure 1 (SEQ ID NO: 2). Decreased activity may be readily measured by one skilled in the art, for example determining the presence of an amino acid variation from the the sequence in Figure 1 (SEQ ID NO: 2) followed by using the aforementioned enzyme binding assay or by measurement mismatch-nicking activity and glycosylase activity. The invention also provides a process for diagnosing a cancer comprising determining from a sample derived from a patient a decreased level of expression of polypeptide having the sequence of Figure 1 (SEQ ID NO: 2). Decresed protein expression can be measured using, on known quantities of protein, the aforementioned enzyme binding

assay or by measurement mismatch-nicking activity and glycosylase activity and comparing these activities to known quantities of non-variant hMYH polypeptide.

The hMYH polypeptides may also be employed in accordance with the present invention by expression of such polypeptides *in vivo*, which is often referred to as "gene therapy."

Thus, for example, cells from a patient may be engineered with a polynucleotide (DNA or RNA) encoding a polypeptide *ex vivo*, with the engineered cells then being provided to a patient to be treated with the polypeptide. Such methods are well-known in the art and are apparent from the teachings herein. For example, cells may be engineered by the use of a retroviral plasmid vector containing RNA encoding a polypeptide of the present invention.

Similarly, cells may be engineered *in vivo* for expression of a polypeptide *in vivo* by, for example, procedures known in the art. For example, a packaging cell is transduced with a retroviral plasmid vector containing RNA encoding a polypeptide of the present invention such that the packaging cell now produces infectious viral particles containing the gene of interest. These producer cells may be administered to a patient for engineering cells *in vivo* and expression of the polypeptide *in vivo*. These and other methods for administering a polypeptide of the present invention by such method should be apparent to those skilled in the art from the teachings of the present invention.

Retroviruses from which the retroviral plasmid vectors hereinabove mentioned may be derived include, but are not limited to, Moloney Murine Leukemia Virus, spleen necrosis virus, retroviruses such as Rous Sarcoma Virus, Harvey Sarcoma Virus, avian leukosis virus, gibbon ape leukemia virus, human immunodeficiency virus, adenovirus, Myeloproliferative Sarcoma Virus, and mammary tumor virus. In one embodiment, the retroviral plasmid vector is derived from Moloney Murine Leukemia Virus.

The vector includes one or more promoters. Suitable promoters which may be employed include, but are not limited to, the retroviral LTR; the SV40 promoter; and

the human cytomegalovirus (CMV) promoter described in Miller, et al., Biotechniques, Vol. 7, No. 9, 980-990 (1989), or any other promoter (e.g., cellular promoters such as eukaryotic cellular promoters including, but not limited to, the histone, pol III, and β-actin promoters). Other viral promoters which may be employed include, but are not limited to, adenovirus promoters, thymidine kinase (TK) promoters, and B19 parvovirus promoters. The selection of a suitable promoter will be apparent to those skilled in the art from the teachings contained herein.

The nucleic acid sequence encoding the polypeptide of the present invention is under the control of a suitable promoter. Suitable promoters which may be employed include, but are not limited to, adenoviral promoters, such as the adenoviral major late promoter; or hetorologous promoters, such as the cytomegalovirus (CMV) promoter; the respiratory syncytial virus (RSV) promoter; inducible promoters, such as the MMT promoter, the metallothionein promoter; heat shock promoters; the albumin promoter; the ApoAI promoter; human globin promoters; viral thymidine kinase promoters, such as the Herpes Simplex thymidine kinase promoter; retroviral LTRs (including the modified retroviral LTRs hereinabove described); the β-actin promoter; and human growth hormone promoters. The promoter also may be the native promoter which controls the gene encoding the polypeptide.

The retroviral plasmid vector is employed to transduce packaging cell lines to form producer cell lines. Examples of packaging cells which may be transfected include, but are not limited to, the PE501, PA317, ψ-2, ψ-AM, PA12, T19-14X, VT-19-17-H2, ψCRE, ψCRIP, GP+E-86, GP+envAm12, and DAN cell lines as described in Miller, Human Gene Therapy, Vol. 1, pgs. 5-14 (1990), which is incorporated herein by reference in its entirety. The vector may transduce the packaging cells through any means known in the art. Such means include, but are not limited to, electroporation, the use of liposomes, and $CaPO_4$ precipitation. In one alternative, the retroviral plasmid vector may be encapsulated into a liposome, or coupled to a lipid, and then administered to a host.

The producer cell line generates infectious retroviral vector particles which include the nucleic acid sequence(s) encoding the polypeptides. Such retroviral vector particles then may be employed, to transduce eukaryotic cells, either *in vitro* or *in vivo*. The transduced eukaryotic cells will express the nucleic acid sequence(s) encoding the polypeptide. Eukaryotic cells which may be transduced include, but are not limited to, embryonic stem cells, embryonic carcinoma cells, as well as hematopoietic stem cells, hepatocytes, fibroblasts, myoblasts, keratinocytes, endothelial cells, and bronchial epithelial cells.

Once the hMYH gene is being expressed intracellularly, it may be employed to repair DNA mismatches and therefore, prevent cells from uncontrolled growth and neoplasia such as occurs in cancer and tumors.

The hMYH gene and gene product of the present invention may be employed to treat patients who have a defect in the hMYH gene. Among the disorders which may be treated in such cases is cancer, and most particularly colon cancer, such as for example HNPCC, as well as xerederma pigmentosum.

hMYH may also be employed to repair oxidative damage to and oxidation of DNA and prevent mutations from oxidative lesions and other modifications of DNA that can be repaired by hMYH. Skilled artisans will be able to use the DNA repair assays of the invention to determine which defects and/or modifications of DNA can be repaired by hMYH.

In accordance with a further aspect of the invention, there is provided a process for determining susceptibility to cancer, and particularly colon cancer, and most particularly HNPCC. Thus, a mutation in hMYH indicates a susceptibility to cancer, and the nucleic acid sequences described above may be employed in an assay for ascertaining such susceptibility. Thus, for example, the assay may be employed to determine a mutation in a human DNA repair protein as herein described, such as a deletion, truncation, insertion, frame shift, etc., with such mutation being indicative of a susceptibility to cancer.

A mutation may be ascertained for example, by a DNA sequencing assay. Tissue samples, including but not limited to blood samples are obtained from a human patient. The samples are processed by methods known in the art to capture the RNA. First strand cDNA is synthesized from the RNA samples by adding an oligonucleotide primer consisting of polythymidine residues which hybridize to the polyadenosine stretch present on the mRNA's. Reverse transcriptase and deoxynucleotides are added to allow synthesis of the first strand cDNA. Primer sequences are synthesized based on the DNA sequence of the DNA repair protein of the invention. The primer sequence is generally comprised of at least 15 consecutive bases, and may contain at least 30 or even 50 consecutive bases.

Individuals carrying mutations in the gene of the present invention may also be detected at the DNA level by a variety of techniques. Nucleic acids for diagnosis may be obtained from a patient's cells, including but not limited to blood, urine, saliva, tissue biopsy and autopsy material. The genomic DNA may be used directly for detection or may be amplified enzymatically by using PCR (Saiki *et al.*, Nature, 324:163-166 (1986)) prior to analysis. RT-PCR can also be used to detect mutations. It is particularly preferred to used RT-PCR in conjunction with automated detection systems, such as, for example, GeneScan. RNA or cDNA may also be used for the same purpose, PCR or RT-PCR. As an example, PCR primers complementary to the nucleic acid encoding hMYH can be used to identify and analyze mutations. Examples of representative primers are shown below in Table 1. For example, deletions and insertions can be detected by a change in size of the amplified product in comparison to the normal genotype. Point mutations can be identified by hybridizing amplified DNA to radiolabeled RNA or alternatively, radiolabeled antisense DNA sequences. Perfectly matched sequences can be distinguished from mismatched duplexes by RNase A digestion or by differences in melting temperatures.

**Table 1**

**Primers used for detection of mutations in hMYH gene**

| | | SEQ ID NO: |
|---|---|---|
| 1 | 5' TCCTCTGAAGCTTGAGGAGCCTCTAGAACT 3' | 10 |
| 2 | 5' TAGCTCCATGGCTGCTTGGTTGAAA 3' | 11 |
| 3 | 5' GCCATCATGAGGAAGCCACGAGCAG 3' | 12 |
| 4 | 5' TAGCTCCATGGCTGCTTGGTTGAAA 3' | 13 |
| 5 | 5' TTGACCCGAAACTGCTGAATAG 3' | 14 |
| 6 | 5' CAGTGGAGATGTGAGACCGAAAGAA 3' | 15 |
| 7 | 5' CAGCCCGGCCAGGAGATTTCAACCA 3' | 16 |
| 8 | 5' CAGTGGAGATGTGAGACCGAAAGAA 3' | 17 |
| 9 | 5' CCCTCACTAAAGGGAACAAAAGCTGG 3' | 18 |

The above primers may be used for amplifying hMYH cDNA isolated from a sample derived from a patient. The invention also provides the primers of Table 1 with 1, 2, 3 or 4 nucleotides removed from the 5' and/or the 3' end. The primers may be used to amplify the gene isolated from the patient such that the gene may then be subject to various techniques for elucidation of the DNA sequence. In this way, mutations in the DNA sequence may be diagnosed.

Sequence differences between the reference gene and genes having mutations may be revealed by the direct DNA sequencing method. In addition, cloned DNA segments may be employed as probes to detect specific DNA segments. The sensitivity of this method is greatly enhanced when combined with PCR. For example, a sequencing primer is used with double-stranded PCR product or a single-stranded template molecule generated by a modified PCR. The sequence determination is performed by conventional procedures with radiolabeled nucleotide or by automatic sequencing procedures with fluorescent-tags.

Genetic testing based on DNA sequence differences may be achieved by detection of alteration in electrophoretic mobility of DNA fragments in gels with or without denaturing agents. Small sequence deletions and insertions can be visualized by high resolution gel electrophoresis. DNA fragments of different sequences may be distinguished on denaturing formamide gradient gels in which the mobilities of different DNA fragments are retarded in the gel at different positions according to their specific melting or partial melting temperatures (see, e.g., Myers *et al.*, Science, 230:1242 (1985)).

Sequence changes at specific locations may also be revealed by nuclease protection assays, such as RNase and S1 protection or the chemical cleavage method (e.g., Cotton *et al.*, PNAS, USA, 85:4397-4401 (1985)).

Thus, the detection of a specific DNA sequence and/or quantitation of the level of the sequence may be achieved by methods such as hybridization, RNase protection, chemical cleavage, direct DNA sequencing or the use of restriction enzymes, (e.g., Restriction Fragment Length Polymorphisms (RFLP)) and Southern blotting of genomic DNA. The invention provides a process for diagnosing, disease, particularly a cancer, and most particularly colon cancer, such as for example HNPCC, comprising determining from a sample derived from a patient a decreased level of expression of polynucleotide having the sequence of Figure 1 (SEQ ID NO: 1). Decreased expression of polynucleotide can be measured using any on of the methods well known in the art for the quantation of polynucleotides, such as, for example, PCR, RT-PCR, RNase protection, Northern blotting and other hybridization methods.

In addition to more conventional gel-electrophoresis and DNA sequencing, mutations can also be detected by *in situ* analysis.

Fluorescence *in situ* hybridization (FISH) of a cDNA clone to a metaphase chromosomal spread can be used to provide a precise chromosomal location.

As an example of how this was performed, hMYH DNA was digested and purified with QIAEX II DNA purification kit (QIAGEN, Inc., Chatsworth, CA) and ligated to Super Cos1 cosmid vector (STRATAGENE, La Jolla, CA). DNA was

purified using Qiagen Plasmid Purification Kit (QIAGEN Inc., Chatsworth, CA) and 1 mg was labeled by nick translation in the presence of Biotin-dATP using BioNick Labeling Kit (GibcoBRL, Life Technologies Inc., Gaithersburg, MD). Biotinilation was detected with GENE-TECT Detection System (CLONTECH Laboratories, Inc. Palo Alto, CA). In situ Hybridization was performed on slides using ONCOR Light Hybridization Kit (ONCOR, Gaithersberg, MD) to detect single copy sequences on metaphase chromosomes. Peripheral blood of normal donors was cultured for three days in RPMI 1640 supplemented with 20% FCS, 3% PHA and penicillin/ streptomycin, synchronized with $10^{-7}$ M methotrexate for 17 hours and washed twice with unsupplemented RPMI. Cells were incubated with $10^{-3}$ M thymidine for 7 hours. The cells were arrested in metaphase after 20 minutes incubation with colcemid (0.5 µg/ml) followed by hypotonic lysis in 75 mM KCl for 15 minutes at 37°C. Cell pellets were then spun out and fixed in Carnoy's fixative (3:1 methanol/acetic acid).

Metaphase spreads were prepared by adding a drop of the suspension onto slides and aid dried. Hybridization was performed by adding 100 ng of probe suspended in 10 ml of hybridization mix (50% formamide, 2xSSC, 1% dextran sulfate) with blocking human placental DNA 1 µg/ml), Probe mixture was denatured for 10 minutes in 70°C water bath and incubated for 1 hour at 37°C, before placing on a prewarmed (37°C) slide, which was previously denatured in 70% formamide/2xSSC at 70°C, and dehydrated in ethanol series, chilled to 4°C.

Slides were incubated for 16 hours at 37°C in a humidified chamber. Slides were washed in 50% formamide/2xSSC for 10 minutes at 41°C and 2xSSC for 7 minutes at 37°C. Hybridization probe was detected by incubation of the slides with FITC-Avidin (ONCOR, Gaithersberg, MD), according to the manufacturer protocol. Chromosomes were counterstained with propridium iodine suspended in mounting medium. Slides were visualized using a Leitz ORTHOPLAN 2-epifluorescence microscope and five computer images were taken using Imagenetics Computer and MacIntosh printer. hMYH maps to the short arm of chromosome 1, between p32.1 and p34.3.

Once a sequence has been mapped to a precise chromosomal location, the physical position of the sequence on the chromosome can be correlated with genetic map data. Such data are found, for example, in V. McKusick, Mendelian Inheritance in Man (publicly available on line via computer). The relationship between genes and diseases that have been mapped to the same chromosomal region are then identified through linkage analysis (Co-Inheritance of Physically Adjacent Genes).

The polypeptides, their fragments or other derivatives, or analogs thereof, or cells expressing them can be used as an immunogen to produce antibodies thereto. These antibodies can be, for example, polyclonal or monoclonal antibodies. The present invention also includes chimeric, single chain, and humanized antibodies, as well as Fab fragments, or the product of an Fab expression library. Various procedures known in the art may be used for the production of such antibodies and fragments.

Antibodies generated against the polypeptides corresponding to a sequence of the present invention can be obtained by direct injection of the polypeptides into an animal or by administering the polypeptides to an animal, preferably a nonhuman. The antibody so obtained will then bind the polypeptides itself. In this manner, even a sequence encoding only a fragment of the polypeptides can be used to generate antibodies binding the whole native polypeptides. Such antibodies can then be used to isolate the polypeptide from tissue expressing that polypeptide.

For preparation of monoclonal antibodies, any technique which provides antibodies produced by continuous cell line cultures can be used. Examples include the hybridoma technique (Kohler and Milstein, 1975, Nature, 256:495-497), the trioma technique, the human B-cell hybridoma technique (Kozbor et al., 1983, Immunology Today 4:72), and the EBV-hybridoma technique to produce human monoclonal antibodies (Cole, et al., 1985, in Monoclonal Antibodies and Cancer Therapy, Alan R. Liss, Inc., pp. 77-96).

Techniques described for the production of single chain antibodies (U.S. Patent 4,946,778) can be adapted to produce single chain antibodies to immunogenic

polypeptide products of this invention. Also, transgenic mice may be used to express humanized antibodies to immunogenic polypeptide products of this invention.

The present invention will be further described with reference to the following examples; however, it is to be understood that the present invention is not limited to such examples. All parts or amounts, unless otherwise specified, are by weight.

In order to facilitate understanding of the following examples certain frequently occurring methods and/or terms will be described.

"Plasmids" are designated by a lower case p preceded and/or followed by capital letters and/or numbers. The starting plasmids herein are either commercially available, publicly available on an unrestricted basis, or can be constructed from available plasmids in accord with published procedures. In addition, equivalent plasmids to those described are known in the art and will be apparent to the ordinarily skilled artisan.

"Digestion" of DNA refers to catalytic cleavage of the DNA with a restriction enzyme that acts only at certain sequences in the DNA. The various restriction enzymes used herein are commercially available and their reaction conditions, cofactors and other requirements were used as would be known to the ordinarily skilled artisan. For analytical purposes, typically 1 μg of plasmid or DNA fragment is used with about 2 units of enzyme in about 20 μl of buffer solution. For the purpose of isolating DNA fragments for plasmid construction, typically 5 to 50 μg of DNA are digested with 20 to 250 units of enzyme in a larger volume. Appropriate buffers and substrate amounts for particular restriction enzymes are specified by the manufacturer. Incubation times of about 1 hour at 37°C are ordinarily used, but may vary in accordance with the supplier's instructions. After digestion the reaction is electrophoresed directly on a polyacrylamide gel to isolate the desired fragment.

Size separation of the cleaved fragments is performed using 8 percent polyacrylamide gel described by Goeddel, D. *et al.*, Nucleic Acids Res., 8:4057 (1980).

"Oligonucleotides" refers to either a single stranded polydeoxynucleotide or two complementary polydeoxynucleotide strands which may be chemically synthesized. Such synthetic oligonucleotides have no 5' phosphate and thus will not ligate to another oligonucleotide without adding a phosphate with an ATP in the presence of a kinase. A synthetic oligonucleotide will ligate to a fragment that has not been dephosphorylated.

"Ligation" refers to the process of forming phosphodiester bonds between two double stranded nucleic acid fragments (Maniatis, T., et al., Id., p. 146). Unless otherwise provided, ligation may be accomplished using known buffers and conditions with 10 units of T4 DNA ligase ("ligase") per 0.5 μg of approximately equimolar amounts of the DNA fragments to be ligated.

Unless otherwise stated, transformation was performed as described in the method of Graham, F. and Van der Eb, A., Virology, 52:456-457 (1973).

## Example 1

**Bacterial Expression and Purification of hMYH**

The DNA sequence encoding hMYH is initially amplified using PCR oligonucleotide primers corresponding to the 5' sequences of the processed hMYH protein (minus the signal peptide sequence) and the vector sequences 3' to the hMYH gene. Additional nucleotides corresponding to hMYH were added to the 5' and 3' sequences respectively. The 5' oligonucleotide primer has the sequence 5' CGC**GGATCC**GCCATCATGACACCGCTCGTCTCC 3' (SEQ ID NO:3) contains a BamHI restriction enzyme site followed by 18 nucleotides of hMYH coding sequence starting from the presumed terminal amino acid of the processed protein codon. The 3' sequence 5' GCG**TCTAGA**TCACTGGGCTGCACTGTTG 3' (SEQ ID NO:4) contains complementary sequences to XbaI site and is followed by 19 nucleotides of hMYH. The restriction enzyme sites correspond to the restriction enzyme sites on the bacterial expression vector pQE-9 (Qiagen, Inc. 9259 Eton Avenue, Chatsworth, CA, 91311)). pQE-9 encodes antibiotic resistance ($Amp^r$), a bacterial origin of replication

(ori), an IPTG-regulatable promoter operator (P/O), a ribosome binding site (RBS), a 6-His tag and restriction enzyme sites. pQE-9 then is digested with BamHI and XbaI. The amplified sequences are ligated into pQE-9 and were inserted in frame with the sequence encoding for the histidine tag and the RBS. The ligation mixture is then used to transform E. coli strain M15/rep 4 (Qiagen, Inc.) by the procedure described in Sambrook, J. et al., Molecular Cloning: A Laboratory Manual, Cold Spring Laboratory Press, (1989). M15/rep4 contains multiple copies of the plasmid pREP4, which expresses the lacI repressor and also confers kanamycin resistance ($Kan^r$). Transformants are identified by their ability to grow on LB plates and ampicillin/kanamycin resistant colonies were selected. Plasmid DNA was isolated and confirmed by restriction analysis. Clones containing the desired constructs are grown overnight (O/N) in liquid culture in LB media supplemented with both Amp (100 ug/ml) and Kan (25 ug/ml). The O/N culture is used to inoculate a large culture at a ratio of 1:100 to 1:250. The cells are grown to an optical density 600 ($O.D._{600}$) of between 0.4 and 0.6. IPTG ("Isopropyl-B-D-thiogalacto pyranoside") is then added to a final concentration of 1 mM. IPTG induces by inactivating the lacI repressor, clearing the P/O leading to increased gene expression. Cells are grown an extra 3 to 4 hours. Cells are then harvested by centrifugation. The cell pellet is solubilized in the chaotropic agent 6 Molar Guanidine HCl. After clarification, solubilized hMYH is purified from this solution by chromatography on a Nickel-Chelate column under conditions that allow for tight binding by proteins containing the 6-His (histidine) tag (Hochuli, E. et al., J. Chromatography 411:177-184 (1984)). hMYH is eluted from the column in 6 molar guanidine HCl pH 5.0 and for the purpose of renaturation adjusted to 3 molar guanidine HCl, 100mM sodium phosphate, 10 mmolar glutathione (reduced) and 2 mmolar glutathione (oxidized). After incubation in this solution for 12 hours the protein was dialyzed to 10 mmolar sodium phosphate.

## Example 2

## Cloning and expression of hMYH using the baculovirus expression system

The DNA sequence encoding the full length hMYH protein was amplified using PCR oligonucleotide primers corresponding to the 5' and 3' sequences of the gene:

The 5' primer has the sequence 5' CGC**GGATCC**CGCAATC<u>ATG</u>ACACCGCTCGTCTCC 3' (SEQ ID NO:5) and contains a BamHI restriction enzyme site (in bold) followed by 18 nucleotides resembling an efficient signal for the initiation of translation in eukaryotic cells (Kozak, M., J. Mol. Biol., 196:947-950 (1987) which is just behind the first 6 nucleotides of the hMYH gene (the initiation codon for translation "ATG" is underlined).

The 3' primer has the sequence 5' GCG**TCTAGA**TCACTGGGCTGCACTGTTG 3' (SEQ ID NO:6) and contains the cleavage site for the restriction endonuclease XbaI and a number of nucleotides complementary to the 3' non-translated sequence of the hMYH gene sufficient for stable hybridization. The amplified sequences were isolated from a 1% agarose gel using a commercially available kit ("Geneclean," BIO 101 Inc., La Jolla, Ca.). The fragment was then digested with the endonucleases BamHI and XbaI and then purified again on a 1% agarose gel. This fragment is designated F2.

The vector pA2 (modification of pVL941 vector, discussed below) is used for the expression of the hMYH protein using the baculovirus expression system (for review see: Summers, M.D. and Smith, G.E. 1987, A manual of methods for baculovirus vectors and insect cell culture procedures, Texas Agricultural Experimental Station Bulletin No. 1555). This expression vector contains the strong polyhedrin promoter of the Autographa californica nuclear polyhedrosis virus (AcMNPV) followed by the recognition sites for the restriction endonucleases BamHI and XbaI. The polyadenylation site of the simian virus SV40 is used for efficient polyadenylation. For an easy selection of recombinant virus the beta-galactosidase

gene from E.coli is inserted in the same orientation as the polyhedrin promoter followed by the polyadenylation signal of the polyhedrin gene. The polyhedrin sequences are flanked at both sides by viral sequences for the cell-mediated homologous recombination of co-transfected wild-type viral DNA. Many other baculovirus vectors could be used in place of pA2 such as pAc373, pVL941, pRG1 and pAcIM1 (Luckow, V.A. and Summers, M.D., Virology, 170:31-39).

The plasmid is digested with the restriction enzymes BamHI and XbaI and then dephosphorylated using calf intestinal phosphatase by procedures known in the art. The DNA is then isolated from a 1% agarose gel using the commercially available kit ("Geneclean" BIO 101 Inc., La Jolla, Ca.). This vector DNA is designated V2.

Fragment F2 and the dephosphorylated plasmid V2 were ligated with T4 DNA ligase. E.coli HB101 cells is then transformed and bacteria identified that contained the plasmid (pBachMYH) with the hMYH gene using the enzymes BamHI and XbaI. The sequence of the cloned fragment is confirmed by DNA sequencing.

5 μg of the plasmid pBachMYH is co-transfected with 1.0 μg of a commercially available linearized baculovirus ("BaculoGold™ baculovirus DNA", Pharmingen, San Diego, CA.) using the lipofection method (Felgner et al. Proc. Natl. Acad. Sci. USA, 84:7413-7417 (1987)).

1μg of BaculoGold™ virus DNA and 5 μg of the plasmid pBachMYH are mixed in a sterile well of a microtiter plate containing 50 μl of serum free Grace's medium (Life Technologies Inc., Gaithersburg, MD). Afterwards 10 μl Lipofectin plus 90 μl Grace's medium are added, mixed and incubated for 15 minutes at room temperature. Then the transfection mixture is added drop-wise to the Sf9 insect cells (ATCC CRL 1711) seeded in a 35 mm tissue culture plate with 1 ml Grace's medium without serum. The plate is rocked back and forth to mix the newly added solution. The plate is then incubated for 5 hours at 27°C. After 5 hours the transfection solution is removed from the plate and 1 ml of Grace's insect medium supplemented with 10% fetal calf serum is added. The plate is put back into an incubator and cultivation continued at 27°C for four days.

After four days the supernatant is collected and a plaque assay performed similar as described by Summers and Smith (supra). As a modification an agarose gel with "Blue Gal" (Life Technologies Inc., Gaithersburg) is used which allows an easy isolation of blue stained plaques. (A detailed description of a "plaque assay" can also be found in the user's guide for insect cell culture and baculovirology distributed by Life Technologies Inc., Gaithersburg, page 9-10).

Four days after the serial dilution, the virus is added to the cells and blue stained plaques are picked with the tip of an Eppendorf pipette. The agar containing the recombinant viruses is then resuspended in an Eppendorf tube containing 200 µl of Grace's medium. The agar is removed by a brief centrifugation and the supernatant containing the recombinant baculovirus is used to infect Sf9 cells seeded in 35 mm dishes. Four days later the supernatants of these culture dishes are harvested and then stored at 4°C.

Sf9 cells are grown in Grace's medium supplemented with 10% heat-inactivated FBS. The cells are infected with the recombinant baculovirus V-hMYH at a multiplicity of infection (MOI) of 2. Six hours later the medium is removed and replaced with SF900 II medium minus methionine and cysteine (Life Technologies Inc., Gaithersburg). 42 hours later 5 µCi of $^{35}$S-methionine and 5 µCi $^{35}$S cysteine (Amersham) are added. The cells are further incubated for 16 hours before they are harvested by centrifugation and the labelled proteins visualized by SDS-PAGE and autoradiography.

## Example 3

## Expression of Recombinant hMYH in COS cells

The expression of plasmid, hMYH HA is derived from a vector pcDNAI/Amp (Invitrogen) containing: 1) SV40 origin of replication, 2) ampicillin resistance gene, 3) E.coli replication origin, 4) CMV promoter followed by a polylinker region, an SV40 intron and polyadenylation site. A DNA fragment encoding the entire hMYH precursor and a HA tag fused in frame to its 3' end was cloned into the polylinker

region of the vector, therefore, the recombinant protein expression is directed under the CMV promoter. The HA tag corresponds to an epitope derived from the influenza hemagglutinin protein as previously described (I. Wilson, H. Niman, R. Heighten, A Cherenson, M. Connolly, and R. Lerner, 1984, Cell 37:767, (1984)). The infusion of HA tag to the target protein allows easy detection of the recombinant protein with an antibody that recognizes the HA epitope.

The plasmid construction strategy is described as follows:

The DNA sequence encoding hMYH is constructed by PCR using two primers: the 5' primer 5'-CGCGGATCCGCCATCATGACACCGCTCGTCTCC-3' (SEQ ID NO:7) contains a BamHI site followed by 18 nucleotides of hMYH coding sequence starting from the initiation codon; the 3' sequence 5'-GCGCTCGAGCTGGGCTGCACTGTTGAGG (SEQ ID NO:8) contains complementary sequences to XhoI site, translation stop codon, HA tag and the last 19 nucleotides of the hMYH coding sequence (not including the stop codon). Therefore, the PCR product contains a BamHI site, hMYH coding sequence and an XhoI site. The PCR amplified DNA fragment and the vector, pcDNAI/Amp (comprising an HA tag at the 3' end), are digested with BamHI and XhoI restriction enzyme and ligated. The ligation mixture is transformed into E. coli strain SURE (available from Stratagene Cloning Systems, La Jolla) the transformed culture is plated on ampicillin media plates and resistant colonies are selected. Plasmid DNA is isolated from transformants and examined by restriction analysis for the presence of the correct fragment. For expression of the recombinant hMYH, COS cells are transfected with the expression vector by DEAE-DEXTRAN method (J. Sambrook, E. Fritsch, T. Maniatis, Molecular Cloning: A Laboratory Manual, Cold Spring Laboratory Press, (1989)). The expression of the hMYH HA protein is detected by radiolabelling and immunoprecipitation method (E. Harlow, D. Lane, Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory Press, (1988)). Cells are labelled for 8 hours with $^{35}$S-cysteine two days post transfection. Culture media is then collected and cells are lysed with detergent (RIPA buffer (150 mM NaCl, 0.1% SDS, 1% NP-40, 0.5% DOC,

50mM Tris, pH 7.5) (Wilson, I. et al., Id. 37:767 (1984)). Both cell lysate and culture media are precipitated with an HA specific monoclonal antibody. Proteins precipitated are analyzed on 15% SDS-PAGE gels.

## Example 4

### Expression via Gene Therapy

Fibroblasts are obtained from a subject by skin biopsy. The resulting tissue is placed in tissue-culture medium and separated into small pieces. Small chunks of the tissue are placed on a wet surface of a tissue culture flask, approximately ten pieces are placed in each flask. The flask is turned upside down, closed tight and left at room temperature over night. After 24 hours at room temperature, the flask is inverted and the chunks of tissue remain fixed to the bottom of the flask and fresh media (e.g., Ham's F12 media, with 10% FBS, penicillin and streptomycin, is added. This is then incubated at 37°C for approximately one week. At this time, fresh media is added and subsequently changed every several days. After an additional two weeks in culture, a monolayer of fibroblasts emerge. The monolayer is trypsinized and scaled into larger flasks.

pMV-7 (Kirschmeier, P.T. et al, DNA, 7:219-25 (1988) flanked by the long terminal repeats of the Moloney murine sarcoma virus, is digested with EcoRI and HindIII and subsequently treated with calf intestinal phosphatase. The linear vector is fractionated on agarose gel and purified, using glass beads.

The cDNA encoding a polypeptide of the present invention is amplified using PCR primers which correspond to the 5' and 3' end sequences respectively. The 5' primer containing an EcoRI site and the 3' primer further includes a HindIII site. Equal quantities of the Moloney murine sarcoma virus linear backbone and the amplified EcoRI and HindIII fragment are added together, in the presence of T4 DNA ligase. The resulting mixture is maintained under conditions appropriate for ligation of the two fragments. The ligation mixture is used to transform bacteria HB101, which

are then plated onto agar-containing kanamycin for the purpose of confirming that the vector had the gene of interest properly inserted.

The amphotropic pA317 or GP+am12 packaging cells are grown in tissue culture to confluent density in Dulbecco's Modified Eagles Medium (DMEM) with 10% calf serum (CS), penicillin and streptomycin. The MSV vector containing the gene is then added to the media and the packaging cells are transduced with the vector. The packaging cells now produce infectious viral particles containing the gene (the packaging cells are now referred to as producer cells).

Fresh media is added to the transduced producer cells, and subsequently, the media is harvested from a 10 cm plate of confluent producer cells. The spent media, containing the infectious viral particles, is filtered through a millipore filter to remove detached producer cells and this media is then used to infect fibroblast cells. Media is removed from a sub-confluent plate of fibroblasts and quickly replaced with the media from the producer cells. This media is removed and replaced with fresh media. If the titer of virus is high, then virtually all fibroblasts will be infected and no selection is required. If the titer is very low, then it is necessary to use a retroviral vector that has a selectable marker, such as neo or his.

The engineered fibroblasts are then injected into the host, either alone or after having been grown to confluence on cytodex 3 microcarrier beads. The fibroblasts now produce the protein product.

Numerous modifications and variations of the present invention are possible in light of the above teachings and, therefore, within the scope of the appended claims, the invention may be practiced otherwise than as particularly described.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: Wei

(ii) TITLE OF INVENTION: Human MutY

(iii) NUMBER OF SEQUENCES: 18

(iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: SMITHKLINE BEECHAM CORPORATION
(B) STREET: 709 SWEDELAND ROAD
(C) CITY: KING OF PRUSSIA
(D) STATE: PENNSYLVANIA
(E) COUNTRY: USA
(F) ZIP: 19406

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: 3.5 INCH DISKETTE
(B) COMPUTER: Compaq LTE Lite 4/33C
(C) OPERATING SYSTEM: MS-DOS
(D) SOFTWARE: MS WORD

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER:
(B) FILING DATE: Herewith
(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:
(A) APPLICATION NUMBER: None
(B) FILING DATE: None

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: GIMMI, EDWARD R.
(B) REGISTRATION NUMBER: 38,891
(C) REFERENCE/DOCKET NUMBER: ATG50002

(viii) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: 610-270-4478
(B) TELEFAX: 610-270-4026

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 1811 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

CTAGTTCAGG CGGAAGGAGC AGTCCTCTGA AGCTTGAGGA GCCTCTAGAA CTATGAGCCC
60
GAGGCCTTCC CCTCTCCCAG AGCGCAGAGG CTTTGAAGGC TACCTCTGGG AAGCCGCTCA
120
CCGTCGGAAG CTGCGGGAGC TGAAACTGCG CCATCGTCAC TGTCGGCGGC C ATG ACA
177 Met Thr

CCG CTC GTC TCC CGC CTG AGT CGT CTG TGG GCC ATC ATG AGG AAG CCA 225
Pro Leu Val Ser Arg Leu Ser Arg Leu Trp Ala Ile Met Arg Lys Pro
5 10 15
CGA GCA GCC GTG GGA AGT GGT CAC AGG AAG CAG GCA GCC AGC CAG GAA 273
Arg Ala Ala Val Gly Ser Gly His Arg Lys Gln Ala Ala Ser Gln Glu
20 25 30
GGG AGG CAG AAG CAT GCT AAG AAC AAC AGT CAG GCC AAG CCT TCT GCC 321
Gly Arg Gln Lys His Ala Lys Asn Asn Ser Gln Ala Lys Pro Ser Ala
35 40 45 50
TGT GAT GGC CTG GCC AGG CAG CCG GAA GAG GTG GTA TTG CAG GCC TCT 369
Cys Asp Gly Leu Ala Arg Gln Pro Glu Glu Val Val Leu Gln Ala Ser
55 60 65
GTC TCC TCA TAC CAT CTA TTC AGA GAC GTA GCT GAA GTC ACA GCC TTC 417
Val Ser Ser Tyr His Leu Phe Arg Asp Val Ala Glu Val Thr Ala Phe
70 75 80
CGA GGG AGC CTG CTA AGC TGG TAC GAC CAA GAG AAA CGG GAC CTA CCA 465
Arg Gly Ser Leu Leu Ser Trp Tyr Asp Gln Glu Lys Arg Asp Leu Pro
85 90 95
TGG AGA AGA CGG GCA GAA GAT GAG ATG GAC CTG GAC AGG CGG GCA TAT 513
Trp Arg Arg Arg Ala Glu Asp Glu Met Asp Leu Asp Arg Arg Ala Tyr
100 105 110
GCT GTG TGG GTC TCA GAG GTC ATG CTG CAG CAG ACC CAG GTT GCC ACT 561
Ala Val Trp Val Ser Glu Val Met Leu Gln Gln Thr Gln Val Ala Thr
115 120 125 130
GTG ATC AAC TAC TAT ACC GGA TGG ATG CAG AAG TGG CCT ACA CTG CAG 609
Val Ile Asn Tyr Tyr Thr Gly Trp Met Gln Lys Trp Pro Thr Leu Gln
135 140 145
GAC CTG GCC AGT GCT TCC CTG GAG GAG GTG AAT CAA CTC TGG GCT GGC 657
Asp Leu Ala Ser Ala Ser Leu Glu Glu Val Asn Gln Leu Trp Ala Gly
150 155 160

CTG GGC TAC TAT TCT CGT GGC CGG CGG CTG CAG GAG GGA GCT CGG AAG 705
Leu Gly Tyr Tyr Ser Arg Gly Arg Arg Leu Gln Glu Gly Ala Arg Lys
165 170 175
GTG GTA GAG GAG CTA GGG GGC CAC ATG CCA CGT ACA GCA GAG ACC CTG 753
Val Val Glu Glu Leu Gly Gly His Met Pro Arg Thr Ala Glu Thr Leu
180 185 190
CAG CAG CTC CTG CCT GGC GTG GGG CGC TAC ACA GCT GGG GCC ATT GCC 801
Gln Gln Leu Leu Pro Gly Val Gly Arg Tyr Thr Ala Gly Ala Ile Ala
195 200 205 210
TCT ATC GCC TTT GGC CAG GCA ACC GGT GTG GTG GAT GGC AAC GTA GCA 849
Ser Ile Ala Phe Gly Gln Ala Thr Gly Val Val Asp Gly Asn Val Ala
215 220 225
CGG GTG CTG TGC CGT GTC CGA GCC ATT GGT GCT GAT CCC AGC AGC ACC 849
Arg Val Leu Cys Arg Val Arg Ala Ile Gly Ala Asp Pro Ser Ser Thr
230 235 240
CTT GTT TCC CAG CAG CTC TGG GGT CTA GCC CAG CAG CTG GTG GAC CCA 897
Leu Val Ser Gln Gln Leu Trp Gly Leu Ala Gln Gln Leu Val Asp Pro
245 250 255
GCC CGG CCA GGA GAT TTC AAC CAA GCA GCC ATG GAG CTA GGG GCC ACA 945
Ala Arg Pro Gly Asp Phe Asn Gln Ala Ala Met Glu Leu Gly Ala Thr
260 265 270
GTG TGT ACC CCA CAG CGC CCA CTG TGC AGC CAG TGC CCT GTG GAG AGC 993
Val Cys Thr Pro Gln Arg Pro Leu Cys Ser Gln Cys Pro Val Glu Ser
275 280 285 290
CTG TGC CGG GCA CGC CAG AGA GTG GAG CAG GAA CAG CTC TTA GCC TCA 1041
Leu Cys Arg Ala Arg Gln Arg Val Glu Gln Glu Gln Leu Leu Ala Ser
295 300 305
GGG AGC CTG TCG GGC AGT CCT GAC GTG GAG GAG TGT GCT CCC AAC ACT 1089
Gly Ser Leu Ser Gly Ser Pro Asp Val Glu Glu Cys Ala Pro Asn Thr
310 315 320
GGA CAG TGC CAC CTG TGC CTG CCT CCC TCG GAG CCC TGG GAC CAG ACC 1137
Gly Gln Cys His Leu Cys Leu Pro Pro Ser Glu Pro Trp Asp Gln Thr
325 330 335
CTG GGA GTG GTC AAC TTC CCC AGA AAG GCC AGC CGC AAG CCC CCC AGG 1185
Leu Gly Val Val Asn Phe Pro Arg Lys Ala Ser Arg Lys Pro Pro Arg
340 345 350
GAG GAG AGC TCT GCC ACC TGT GTT CTG GAA CAG CCT GGG GCC CTT GGG 1233
Glu Glu Ser Ser Ala Thr Cys Val Leu Glu Gln Pro Gly Ala Leu Gly
355 360 365 370
GCC CAA ATT CTG CTG GTG CAG AGG CCC AAC TCA GGT CTG CTG GCA GGA 1281
Ala Gln Ile Leu Leu Val Gln Arg Pro Asn Ser Gly Leu Leu Ala Gly
375 380 385
CTG TGG GAG TTC CCG TCC GTG ACC TGG GAG CCC TCA GAG CAG CTT CAG 1329
Leu Trp Glu Phe Pro Ser Val Thr Trp Glu Pro Ser Glu Gln Leu Gln
390 395 400
CGC AAG GCC CTG CTG CAG GAA CTA CAG CGT TGG GCT GGG CCC CTC CCA 1377
Arg Lys Ala Leu Leu Gln Glu Leu Gln Arg Trp Ala Gly Pro Leu Pro
405 410 415
GCC ACG CAC CTC CGG CAC CTT GGG GAG GTT GTC CAC ACC TTC TCT CAC 1425
Ala Thr His Leu Arg His Leu Gly Glu Val Val His Thr Phe Ser His
420 425 430
ATC AAG CTG ACA TAT CAA GTA TAT GGG CTG GCC TTG GAA GGG CAG ACC 1473

Ile Lys Leu Thr Tyr Gln Val Tyr Gly Leu Ala Leu Glu Gly Gln Thr
435 440 445 450
CCA GTG ACC ACC GTA CCA CCA GGT GCT CGC TGG CTG ACG CAG GAG GAA 1521
Pro Val Thr Thr Val Pro Pro Gly Ala Arg Trp Leu Thr Gln Glu Glu
455 460 465
TTT CAC ACC GCA GCT GTT TCC ACC GCC ATG AAA AAG GTT TTC CGT GTG 1569
Phe His Thr Ala Ala Val Ser Thr Ala Met Lys Lys Val Phe Arg Val
470 475 480
TAT CAG GGC CAA CAG CCA GGG ACC TGT ATG GGT TCC AAA AGG TCC CAG 1617
Tyr Gln Gly Gln Gln Pro Gly Thr Cys Met Gly Ser Lys Arg Ser Gln
485 490 495
GTG TCC TCT CCG TGC AGT CGG AAA AAG CCC CGC ATG GGC CAG CAA GTC 1665
Val Ser Ser Pro Cys Ser Arg Lys Lys Pro Arg Met Gly Gln Gln Val
500 505 510
CTG GAT AAT TTC TTT CGG TCT CAC ATC TCC ACT GAT GCA CAC AGC CTC 1713
Leu Asp Asn Phe Phe Arg Ser His Ile Ser Thr Asp Ala His Ser Leu
515 520 525 530
AAC AGT GCA GCC CAG TGA CACCTCTGAA AGCCCCCATT CCCTGAGAAT 1761
Asn Ser Ala Ala Gln *
535
CCTGTTGTTA GTAAAGTGCT TATTTTTGTA GTTAAAAAAA AAAAAAAAA 1811

(2) INFORMATION FOR SEQ ID NO:2:
(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 535 AMINO ACIDS
(B) TYPE: AMINO ACID
(C) STRANDEDNESS:
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: PROTEIN

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:
Met Thr Pro Leu Val Ser Arg Leu Ser Arg Leu Trp Ala Ile Met
5 10 15
Arg Lys Pro Arg Ala Ala Val Gly Ser Gly His Arg Lys Gln Ala
20 25 30
Ala Ser Gln Glu Gly Arg Gln Lys His Ala Lys Asn Asn Ser Gln
35 40 45
Ala Lys Pro Ser Ala Cys Asp Gly Leu Ala Arg Gln Pro Glu Glu
50 55 60
Val Val Leu Gln Ala Ser Val Ser Ser Tyr His Leu Phe Arg Asp
65 70 75
Val Ala Glu Val Thr Ala Phe Arg Gly Ser Leu Leu Ser Trp Tyr
80 85 90
Asp Gln Glu Lys Arg Asp Leu Pro Trp Arg Arg Arg Ala Glu Asp
95 100 105
Glu Met Asp Leu Asp Arg Arg Ala Tyr Ala Val Trp Val Ser Glu
110 115 120
Val Met Leu Gln Gln Thr Gln Val Ala Thr Val Ile Asn Tyr Tyr

125 130 135

Thr Gly Trp Met Gln Lys Trp Pro Thr Leu Gln Asp Leu Ala Ser

140 145 150

Ala Ser Leu Glu Glu Val Asn Gln Leu Trp Ala Gly Leu Gly Tyr

155 160 165

Tyr Ser Arg Gly Arg Arg Leu Gln Glu Gly Ala Arg Lys Val Val

170 175 180

Glu Glu Leu Gly Gly His Met Pro Arg Thr Ala Glu Thr Leu Gln

185 190 195

Gln Leu Leu Pro Gly Val Gly Arg Tyr Thr Ala Gly Ala Ile Ala

200 205 210

Ser Ile Ala Phe Gly Gln Ala Thr Gly Val Val Asp Gly Asn Val

215 220 225

Ala Arg Val Leu Cys Arg Val Arg Ala Ile Gly Ala Asp Pro Ser

230 235 240

Ser Thr Leu Val Ser Gln Gln Leu Trp Gly Leu Ala Gln Gln Leu

245 250 255

Val Asp Pro Ala Arg Pro Gly Asp Phe Asn Gln Ala Ala Met Glu

260 265 270

Leu Gly Ala Thr Val Cys Thr Pro Gln Arg Pro Leu Cys Ser Gln

275 280 285

Cys Pro Val Glu Ser Leu Cys Arg Ala Arg Gln Arg Val Glu Gln

290 295 300

Glu Gln Leu Leu Ala Ser Gly Ser Leu Ser Gly Ser Pro Asp Val

305 310 315

Glu Glu Cys Ala Pro Asn Thr Gly Gln Cys His Leu Cys Leu Pro

320 325 330

Pro Ser Glu Pro Trp Asp Gln Thr Leu Gly Val Val Asn Phe Pro

335 340 345

Arg Lys Ala Ser Arg Lys Pro Pro Arg Glu Glu Ser Ser Ala Thr

350 355 360

Cys Val Leu Glu Gln Pro Gly Ala Leu Gly Ala Gln Ile Leu Leu

365 370 375

Val Gln Arg Pro Asn Ser Gly Leu Leu Ala Gly Leu Trp Glu Phe

380 385 390

Pro Ser Val Thr Trp Glu Pro Ser Glu Gln Leu Gln Arg Lys Ala

395 400 405

Leu Leu Gln Glu Leu Gln Arg Trp Ala Gly Pro Leu Pro Ala Thr

410 415 420

His Leu Arg His Leu Gly Glu Val Val His Thr Phe Ser His Ile

425 430 435

Lys Leu Thr Tyr Gln Val Tyr Gly Leu Ala Leu Glu Gly Gln Thr

440 445 450

Pro Val Thr Thr Val Pro Pro Gly Ala Arg Trp Leu Thr Gln Glu

455 460 465

Glu Phe His Thr Ala Ala Val Ser Thr Ala Met Lys Lys Val Phe

470 475 480

Arg Val Tyr Gln Gly Gln Gln Pro Gly Thr Cys Met Gly Ser Lys

485 490 495

Arg Ser Gln Val Ser Ser Pro Cys Ser Arg Lys Lys Pro Arg Met

500 505 510

Gly Gln Gln Val Leu Asp Asn Phe Phe Arg Ser His Ile Ser Thr

515 520 525
Asp Ala His Ser Leu Asn Ser Ala Ala Gln
530 535

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 34 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

CGCGGATCCG CCATCATTGA CACCGCTCGT CTCC 34

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 28 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

GCGTCTAGAT CACTGGGCTG CACTGTTG 28

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 34 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

CGCGGATCCC GCAATCATGA CACCGCTCGT CTCC 34

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 28 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

GCGTCTAGAT CACTGGGCTG CACTGTTG 28

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS

(A) LENGTH: 355 AMINO ACIDS

(B) TYPE: AMINO ACID

(C) STRANDEDNESS:

(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: PROTEIN

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

```
Met Gln Ala Ser Gln Phe Ser Ala Gln Val Leu Asp Trp Tyr Asp
                5                  10                  15
Lys Tyr Gly Arg Lys Thr Leu Pro Trp Gln Ile Asp Lys Thr Pro
                20                 25                  30
Tyr Lys Val Trp Leu Se Glu Val Met Leu Gln Gln Thr Gln Val
                35                 40                  45
Ala Thr Val Ile Pro Tyr Phe Glu Arg Phe Met Ala Arg Phe Pro
                50                 55                  60
Thr Val Thr Asp Leu Ala Asn Ala Pro Leu Asp Glu Val Leu His
                65                 70                  75
Leu Trp Thr Gly Leu Gly Tyr Tyr Ala Arg Ala Arg Asn Leu His
                80                 85                  90
Lys Ala Ala Gln Gln Val Ala Thr Leu His Gly Gly Lys Phe Pro
                95                 100                 105
Glu Thr Phe Glu Glu Val Ala Ala Leu Pro Gly Val Gly Arg Ser
                110                115                 120
Thr Ala Gly Ala Ile Leu Ser Leu Ser Leu Gly Lys His Phe Pro
                125                130                 135
Ile Leu Asp Gly Asn Val Lys Arg Val Leu Ala Arg Cys Tyr Ala
                140                145                 150
Val Ser Gly Trp Pro Gly Lys Lys Glu Val Glu Asn Lys Leu Trp
                155                160                 165
Ser Leu Ser Glu Gln Val Thr Pro Ala Val Gly Val Glu Arg Phe
                170                175                 180
Asn Gln Ala Met Met Asp Leu Gly Ala Met Ile Cys Thr Arg Ser
                185                190                 195
Lys Pro Lys Cys Ser Leu Cys Pro Leu Gln Asn Gly Cys Ile Ala
                200                205                 210
Ala Ala Asn Asn Ser Trp Ala Leu Tyr Pro Gly Lys Lys Pro Lys
                215                220                 225
Gln Thr Leu Pro Glu Arg Thr Gly Tyr Phe Leu Leu Leu Gln His
                230                235                 240
Glu Asp Glu Val Leu Leu Ala Gln Arg Pro Pro Ser Gly Leu Trp
                245                250                 255
Gly Gly Leu Tyr Cys Phe Pro Gln Phe Ala Asp Glu Glu Ser Leu
```

```
                260                 265                 270
Arg Gln Trp Leu Ala Gln Arg Gln Ile Ala Ala Asp Asn Leu Thr
                275                 280                 285
Gln Leu Thr Ala Phe Arg His Thr Phe Ser His Phe His Leu Asp
                290                 295                 300
Ile Val Pro Met Trp Leu Pro Val Ser Ser Phe Thr Gly Cys Met
                305                 310                 315
Asp Glu Gly Asn Ala Leu Trp Tyr Asn Leu Ala Gln Pro Pro Ser
                320                 325                 330
Val Gly Leu Ala Ala Pro Val Glu Arg Leu Leu Gln Gln Leu Arg
                335                 340                 350
Thr Gly Ala Pro Val
                355
```

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 30 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

TCCTCTGAAG CTTGAGGAGC CTCTAGAACT 30

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 25 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

TAGCTCCATG GCTGCTTGGT TGAAA 25

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 25 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

GCCATCATGA GGAAGCCACG AGCAG 25

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 25 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

TAGCTCCATG GCTGCTTGGT TGAAA 25

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 22 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

TTGACCCGAA ACTGCTGAAT AG 22

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 25 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

CAGTGGAGAT GTGAGACCGA AAGAA 25

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 25 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

CAGCCCGGCC AGGAGATTTC AACCA 25

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 25 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

CAGTGGAGAT GTGAGACCGA AAGAA 25

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS
(A) LENGTH: 26 BASE PAIRS
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: Oligonucleotide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

CCCTCACTAA AGGGAACAAA AGCTGG 26

WHAT IS CLAIMED IS:

1. An isolated polynucleotide comprising a polynucleotide which is at least 70% identical to a member selected from the group consisting of:

(a) a polynucleotide encoding a polypeptide comprising amino acid 1 to amino acid 535 set forth in SEQ ID NO:2;

(b) a polynucleotide which is complementary to the polynucleotide of (b); and

(c) a polynucleotide comprising at least 30 consecutive bases of the polynucleotide of (a) or (b).

2. The polynucleotide of Claim 1 wherein the polynucleotide is DNA.

3. The polynucleotide of Claim 1 wherein the polynucleotide is RNA.

4. The polynucleotide of Claim 1 wherein the polynucleotide is genomic DNA.

5. The polynucleotide of Claim 2 which encodes the polypeptide comprising amino acid 1 to 535 of SEQ ID NO:2.

6. The polynucleotide of claim 2 comprising the sequence as set forth in SEQ ID NO:1 from nucleotide 1 to nucleotide 1811.

7. The polynucleotide of claim 2 comprising the sequence as set forth in SEQ ID No. 1 from nucleotide 172 to nucleotide 1729.

8. An isolated polynucleotide comprising a polynucleotide having at least a 70% identity to a member selected from the group consisting of:

(a) a polynucleotide which is complementary to the polynucleotide of (a); and

(b) a polynucleotide comprising at least 30 consecutives bases of the polynucleotide of (a).

9. The isolated polynucleotide of claim 8 wherein said polynucleotide is the polynucleotide which expresses hMYH.

10. An isolated polynucleotide comprising a member selected from the group consisting of:

(a) DNA having at least 15 consecutive bases and which is at least 70% complementary to a member selected from the group consisting of:

(i) DNA comprising at least 15 consecutive bases selected from the group consisting of nucleotide 1 to nucleotide 1811 of SEQ ID NO:1;

(ii) DNA complementary to (i); and

(b) RNA corresponding to the DNA of (a).

11. A vector comprising the DNA of Claim 2.

12. A host cell genetically engineered with the vector of Claim 11.

13. A process for using the host cell of Claim 12 comprising: expressing from the host cell a polypeptide encoded by DNA contained in the vector.

14. A process for using a cell comprising: genetically engineering the cell with the vector of claim 11 such that the cell expresses a polypeptide encoded by the DNA contained in said vector.

15. A polypeptide comprising a member selected from the group consisting of:

(a) a polypeptide having an amino acid sequence set forth in SEQ ID NO:2; and

(b) a polypeptide which is at least 70% identical to the polypeptide of (a).

16. The polypeptide of Claim 15 wherein the polypeptide comprises amino acid 1 to amino acid 535 of SEQ ID NO:2.

17. A method for the treatment of a patient having need of hMYH comprising: administering to the patient a therapeutically effective amount of the polypeptide of claim 15 by providing to the patient DNA encoding said polypeptide and expressing said polypeptide *in vivo*.

18. A diagnostic process comprising:
analyzing for the presence of the polypeptide of claim 15 in a sample derived from a host.

19. A process for diagnosing a susceptibility to cancer comprising:
determining from a sample derived from a patient a mutation in the polynucleotide sequence of claim 1.

20. A process for diagnosing a cancer comprising:
determining from a sample derived from a patient a decreased level of activity of a polypeptide having the sequence of claim 15.

21. A process for diagnosing a cancer comprising:
determining from a sample derived from a patient a decreased level of expression of a polypeptide having the sequence of claim 15.

22. A process for diagnosing a cancer comprising:
determining from a sample derived from a patient a decreased level of expression of a polynucleotide having the sequence of claim 1.

23. A process for diagnosing cancer comprising:

determining from a sample derived from a patient a mutation in the polynucleotide sequence of claim 1.

24. An antibody against a polypeptide of claim 15.

```
TCTCCTCG TGGXC TAGTT CAGGC GGAAG GAGCA GTCCT CTGAA GCTTG -183
AGGAG  CCTCT  AGAAC  TATGA GCCCG AGGCC TTCCC CTCTC CCAGA -135
GCGCA  GAGGC  TTTGA  AGGCT ACCTC TGGGA AGCCG CTCAC CGTCG  -90
GAAGC  TGCGG  GAGCT  GAAAC TGCGC CATCG TCACT GTCGG CGGCC  -45

ATG ACA CCG CTC GTC TCC CGC CTG AGT CGT CTG TGG GCC ATC   42
 M   T   P   L   V   S   R   L   S   R   L   W   A   I    14

ATG AGG AAG CCA CGA GCA GCC GTG GGA AGT GGT CAC AGG AAG   84
 M   R   K   P   R   A   A   V   G   S   G   H   R   K    28

CAG GCA GCC AGC CAG GAA GGG AGG CAG AAG CAT GCT AAG AAC  126
 Q   A   A   S   Q   E   G   R   Q   K   H   A   K   N    42

AAC AGT CAG GCC AAG CCT TCT GCC TGT GAT GGC CTG GCC AGG  168
 N   S   Q   A   K   P   S   A   C   D   G   L   A   R    56

CAG CCG GAA GAG GTG GTA TTG CAG GCC TCT GTC TCC TCA TAC  210
 Q   P   E   E   V   V   L   Q   A   S   V   S   S   Y    70

CAT CTA TTC AGA GAC GTA GCT GAA GTC ACA GCC TTC CGA GGG  252
 H   L   F   R   D   V   A   E   V   T   A   F   R   G    84

AGC CTG CTA AGC TGG TAC GAC CAA GAG AAA CGG GAC CTA CCA  294
 S   L   L   S   W   Y   D   Q   E   K   R   D   L   P    98

TGG AGA AGA CGG GCA GAA GAT GAG ATG GAC CTG GAC AGG CGG  336
 W   R   R   R   A   E   D   E   M   D   L   D   R   R   112

GCA TAT GCT GTG TGG GTC TCA GAG GTC ATG CTG CAG CAG ACC  378
 A   Y   A   V   W   V   S   E   V   M   L   Q   Q   T   126

CAG GTT GCC ACT GTG ATC AAC TAC TAT ACC GGA TGG ATG CAG  420
 Q   V   A   T   V   I   N   Y   Y   T   G   W   M   Q   140

AAG TGG CCT ACA CTG CAG GAC CTG GCC AGT GCT TCC CTG GAG  462
 K   W   P   T   L   Q   D   L   A   S   A   S   L   E   154

GAG GTG AAT CAA CTC TGG GCT GGC CTG GGC TAC TAT TCT CGT  504
 E   V   N   Q   L   W   A   G   L   G   Y   Y   S   R   168

GGC CGG CGG CTG CAG GAG GGA GCT CGG AAG GTG GTA GAG GAG  546
 G   R   R   L   Q   E   G   A   R   K   V   V   E   E   182

CTA GGG GGC CAC ATG CCA CGT ACA GCA GAG ACC CTG CAG CAG  588
 L   G   G   H   M   P   R   T   A   E   T   L   Q   Q   196

CTC CTG CCT GGC GTG GGG CGC TAC ACA GCT GGG GCC ATT GCC  630
 L   L   P   G   V   G   R   Y   T   A   G   A   I   A   210

TCT ATC GCC TTT GGC CAG GCA ACC GGT GTG GTG GAT GGC AAC  672
 S   I   A   F   G   Q   A   T   G   V   V   D   G   N   224

GTA GCA CGG GTG CTG TGC CGT GTC CGA GCC ATT GGT GCT GAT  714
 V   A   R   V   L   C   R   V   R   A   I   G   A   D   238

CCC AGC AGC ACC CTT GTT TCC CAG CAG CTC TGG GGT CTA GCC  756
 P   S   S   T   L   V   S   Q   Q   L   W   G   L   A   252

CAG CAG CTG GTG GAC CCA GCC CGG CCA GGA GAT TTC AAC CAA  798
 Q   Q   L   V   D   P   A   R   P   G   D   F   N   Q   266
```

```
GCA GCC ATG GAG CTA GGG GCC ACA GTG TGT ACC CCA CAG CGC   840
 A   A   M   E   L   G   A   T   V   C   T   P   Q   R    280

CCA CTG TGC AGC CAG TGC CCT GTG GAG AGC CTG TGC CGG GCA   882
 P   L   C   S   Q   C   P   V   E   S   L   C   R   A    294

CGC CAG AGA GTG GAG CAG GAA CAG CTC TTA GCC TCA GGG AGC   924
 R   Q   R   V   E   Q   E   Q   L   L   A   S   G   S    308

CTG TCG GGC AGT CCT GAC GTG GAG GAG TGT GCT CCC AAC ACT   966
 L   S   G   S   P   D   V   E   E   C   A   P   N   T    322

GGA CAG TGC CAC CTG TGC CTG CCT CCC TCG GAG CCC TGG GAC  1008
 G   Q   C   H   L   C   L   P   P   S   E   P   W   D    336

CAG ACC CTG GGA GTG GTC AAC TTC CCC AGA AAG GCC AGC CGC  1050
 Q   T   L   G   V   V   N   F   P   R   K   A   S   R    350

AAG CCC CCC AGG GAG GAG AGC TCT GCC ACC TGT GTT CTG GAA  1092
 K   P   P   R   E   E   S   S   A   T   C   V   L   E    364

CAG CCT GGG GCC CTT GGG GCC CAA ATT CTG CTG GTG CAG AGG  1134
 Q   P   G   A   L   G   A   Q   I   L   L   V   Q   R    378

CCC AAC TCA GGT CTG CTG GCA GGA CTG TGG GAG TTC CCG TCC  1176
 P   N   S   G   L   L   A   G   L   W   E   F   P   S    392

GTG ACC TGG GAG CCC TCA GAG CAG CTT CAG CGC AAG GCC CTG  1218
 V   T   W   E   P   S   E   Q   L   Q   R   K   A   L    406

CTG CAG GAA CTA CAG CGT TGG GCT GGG CCC CTC CCA GCC ACG  1260
 L   Q   E   L   Q   R   W   A   G   P   L   P   A   T    420

CAC CTC CGG CAC CTT GGG GAG GTT GTC CAC ACC TTC TCT CAC  1302
 H   L   R   H   L   G   E   V   V   H   T   F   S   H    434

ATC AAG CTG ACA TAT CAA GTA TAT GGG CTG GCC TTG GAA GGG  1344
 I   K   L   T   Y   Q   V   Y   G   L   A   L   E   G    448

CAG ACC CCA GTG ACC ACC GTA CCA CCA GGT GCT CGC TGG CTG  1386
 Q   T   P   V   T   T   V   P   P   G   A   R   W   L    462

ACG CAG GAG GAA TTT CAC ACC GCA GCT GTT TCC ACC GCC ATG  1428
 T   Q   E   E   F   H   T   A   A   V   S   T   A   M    476

AAA AAG GTT TTC CGT GTG TAT CAG GGC CAA CAG CCA GGG ACC  1470
 K   K   V   F   R   V   Y   Q   G   Q   Q   P   G   T    490

TGT ATG GGT TCC AAA AGG TCC CAG GTG TCC TCT CCG TGC AGT  1512
 C   M   G   S   K   R   S   Q   V   S   S   P   C   S    504

CGG AAA AAG CCC CGC ATG GGC CAG CAA GTC CTG GAT AAT TTC  1554
 R   K   K   P   R   M   G   Q   Q   V   L   D   N   F    518

TTT CGG TCT CAC ATC TCC ACT GAT GCA CAC AGC CTC AAC AGT  1596
 F   R   S   H   I   S   T   D   A   H   S   L   N   S    532

GCA GCC CAG TGA TGACA CCTCT GAAAG CCCCC ATTCC CTGAG AATC 1643
 A   A   Q   .                                            535

CTGTTGTT AGTAAA GTGCTT ATTTTT GTAGTT AAAAAA AAAA AAAAAA  1687
```

Figure 2

```
Human   63 LQASVSSYHLFRDVAEVTAFRGSLLSWYDQ.EKRDLPWRRRAEDEMDLDR 111
           :|||                .|.: :|.|||. :::.|||       ::|:
E.coli   1 MQAS..............QFSAQVLDWYDKYGRKTLPW........QIDK 28

       112 RAYAVWVSEVMLQQTQVATVINYYTGWMQKWPTLQDLASASLEEVNQLWA 161
            :| ||:|||||||||||||| |:. :|.::||: |||.|.|:|| :||.
        29 TPYKVWLSEVMLQQTQVATVIPYFERFMARFPTVTDLANAPLDEVLHLWT 78

       162 GLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIAS 211
           |||||.|:|.|::|..|..  ||.:| | |.:.. |||||| ||||| |
        79 GLGYYARARNLHKAAQQVATLHGGKFPETFEEVAA.LPGVGRSTAGAILS 127

       212 IAFGQATGVVDGNVARVLCRVRAIGADPSSTLVSQQLWGLAQQLVDPARP 261
           :.:|.  .::|||| |||.|. |::: |:.. |...||:|.:|:..:. .
       128 LSLGKHFPILDGNVKRVLARCYAVSGWPGKKEVENKLWSLSEQVTPAVGV 177

       262 GDFNQAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSLSG 311
           : |||| |:||| :||. :| || ||::. | |
       178 ERFNQAMMDLGAMICTRSKPKCSLCPLQNGCIA................. 210

       312 SPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATC 361
                  |:|.:   :.| |:.. .|||.             |..:
       211 .......AANNS...WALYPGKKPKQTLP..............ERTGYF 235

       362 VLEQPGALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQELQ 411
           :| |.    :.::||.||| |||::||: || .. |.|       | | |.
       236 LLLQH...EDEVLLAQRPPSGLWGGLYCFPQFADEES.......LRQWLA 275

       412 RWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARW 461
           .:   .::|.:| :|... |||||:.|.. .   |::.: |.. . ...| |
       276 QR..QIAADNLTQLTAFRHTFSHFHLDIVPMWLPVSSFTGCMD.EGNALW 322

       462 LTQEEFHTAAVSTAMKKVFRVYQGQQPG 489
           .. .: ...::.::.:::. .....|.
       323 YNLAQPPSVGLAAPVERLLQQLRTGAPV 350
```

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/03239

**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(6) :C07H 21/04; C12N 15/00, 1/20, 9/14; A61K 38/46
US CL : 536/23.2; 435/320.1, 252.3, 195; 424/94.6
According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 536/23.2; 435/320.1, 252.3, 195; 424/94.6

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Extra Sheet.

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X<br>------<br>Y | J. Cellu. Biochem., Supp. 21A, 10 March 1995, pp.298, left cloumn, the abstract No. C5-240, Radany, E. H. 'Expression Cloning of Candidate Human MutY Homolog.' | 1-16<br>-------<br>17 |
| X<br>-----<br>Y | McGoldrick et al. Characterization of a Mammalian Homolog of the Escherichia coli MutY Mismatch Repair Protein. Mol. and Cellu. Biol. February 1995, Vol. 15, No. 12, pages 989-996. | 15, 16<br>----------<br>1-14, 17 |
| X<br>-----<br>Y | Bessho et al. Evidence for Two DNA Repair Enzymes for 8-Hydroxyguanine (7,8-Dihydro-8-oxoguanine) in Human Cells. J. Biol. Chem. 15 September 1993, Vol. 268, No. 26, pages 19416-19421. | 15, 16<br>----------<br>1-14, 17 |

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

| | |
|---|---|
| Date of the actual completion of the international search<br>19 JUNE 1996 | Date of mailing of the international search report<br>19 JUL 1996 |
| Name and mailing address of the ISA/US<br>Commissioner of Patents and Trademarks<br>Box PCT<br>Washington, D.C. 20231<br>Facsimile No. (703) 305-3230 | Authorized officer<br>NASHAAT T. NASHED<br>Telephone No. (703) 308-0196 |

Form PCT/ISA/210 (second sheet)(July 1992)*

## INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/03239

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A | Michaels et al. The GO System Protects Organisms from the Mutagenic Effect of the Spontaneous Lesion 8-Hydroxyguanine (7,8-Dihydro-8-oxoguanine). J. Bacteriology. October 1992, Vol. 174, No. 20, pages 6321-6325. | 1-17 |
| A | Michaels et al. Evidence that MutY and MutM combine to prevent mutations by an oxidatively damaged form of guanine in DNA. Proc. Natl. Acad. Sci. USA. August 1992, Vol. 89, pages 7022-7025. | 1-17 |

**Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)**

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

**Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)**

This International Searching Authority found multiple inventions in this international application, as follows:

Please See Extra Sheet.

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-17

**Remark on Protest** ☐ The additional search fees were accompanied by the applicant's protest.
☐ No protest accompanied the payment of additional search fees.

B. FIELDS SEARCHED

Electronic data bases consulted (Name of data base and where practicable terms used):

STN Files: Medline, Caplus, Scisearch, Lifesci, Biosis, Embase, Wpids
APS
Search Term: 7,8-dihydro-8-oxoguanine, 8-hydroxyguanine, DNA repair, Muty, Mut y and mammalian not human.

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION WAS LACKING

This ISA found multiple inventions as follows:

I. Claims 1-17, drawn to a polynucleotide sequence encoding for hMYH peptide, a vector containing the polynucleotide, process to make the peptide, the peptide, and a method of treating a patent having need for hMYH.

II. Claims 18, 20 and 21, drawn to diagnostic process in which the presence, activity and level of expression of hMYH are determined.

III. Claims 19, 22, and 23, drawn to diagnostic method in which mutation in and level of expression of the polynucleotide sequence are determined.

IV. Claim 24, drawn to antibody that binds the hMYH peptide.

The inventions listed as Groups I-IV do not relate to a single inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons: Group I comprises a novel gene encoding for a novel protein and methods of making and using the protein. The special technical feature in this Group is the novel protein which is different from those in Groups II-IV. The special technical feature in the method of Group II which detect the hMYH protein is different from that of Group III which assay for sequence differences and level of expression of nucleic acids. Group IV which is antibody is a different and distinct product from the hMYH protein and its nucleic acid (Group I) and the methods of Group II and III. Thus, the claims are not so linked by a special technical feature within the meaning of PCT Rule 13.1 so as to form a single invention concept.